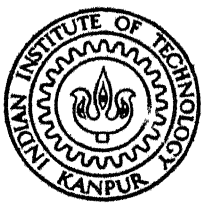
TESTS FOR TWO OR MORE OUTLIERS IN A LINEAR MODEL

By

S. LALITHA



DEPARTMENT OF MATHEMATICS

INDIAN INSTITUTE OF TECHNOLOGY KANPUR

SEPTEMBER, 1983

TESTS FOR TWO OR MORE OUTLIERS IN A LINEAR MODEL

A Thesis Submitted
in Partial Fulfilment of the Requirements
for the Degree of

DOCTOR OF PHILOSOPHY

By
S. LALITHA

to the

DEPARTMENT OF MATHEMATICS
INDIAN INSTITUTE OF TECHNOLOGY KANPUR
SEPTEMBER, 1983

CERTIFICATE

This is to certify that the matter embodied in the thesis entitled "Tests for Two or More Outliers in a Linear Model" by Ms. S. Lalitha for the award of the Degree of Doctor of Philosophy of the Indian Institute of Technology, Kanpur, is a record of bonafide research work carried out by her under my supervision and guidance. The results embodied in this thesis have not been submitted to any other University or Institute for the award of any degree or diploma.

September - 1983

P.C. Joshi 28-9-1983

(P.C. Joshi)
Department of Mathematics
Indian Institute of Technology
KANPUR

CERTIFICATE

This is to certify that Ms. S. Lalitha has satisfactorily completed all the course requirement for the Ph.D. programme in Statistics. The courses include :

M 502	Computer Programming
M 592	Numerical Analysis
M 601	Graduate Mathematics I
M 603	Graduate Mathematics III
M 841	Topics in Statistics
M 843	Topics in Statistical Inference

Ms. S. Lalitha was admitted to the candidacy of the Ph.D. degree in April, 1981 after she successfully completed the written and oral qualifying examinations.

J.B. Shukla) 28/9/83
Head
Department of Mathematics

(Punyatma Singh)
Convener
Departmental Post-graduate
Committee

DEDICATED
TO
MY PARENTS
AND
THESIS SUPERVISOR

ACKNOWLEDGEMENTS

I wish to express my deep sense of gratitude and thanks to my supervisor Professor Prakash C. Joshi for his dynamic guidance and inspiring discussions throughout the research.

I would like to acknowledge the help and inspiration which I had from Professor J.B. Shukla, Head of Mathematics Department, Indian Institute of Technology, Kanpur. I also wish to thank Professors Ishwari D. Dhariyal, Divakar Sharma and all other members of the faculty of the Department of Mathematics for offering stimulating courses and valuable advices.

I am more than grateful to my friends Mr. B.Narayana Shetty, Ms. Sangeeta Khare and Ms. Jayshree Paul for their help in the onerous task of proof reading, without which it would have been difficult to bring out this thesis in the present form.

Finally, special thanks are due to Shri G.L. Misra and A.N. Upadhyaya for their patient typing and cyclostyling the manuscript efficiently.

September - 1983. S. Lalitha

TABLE OF CONTENTS

TABLE OF CONTENTS (continued)

LIST OF TABLES

LIST OF TABLES (Contd.)

LIST OF TABLES (Contd.)

LIST OF TABLES (Contd.)

LIST OF FIGURES

SYNOPSIS

In this thesis we have considered the problem of detection and identification of multiple outliers in a fixed effects linear model. An outlier is an observation that deviates from the rest of the observations in some sense. Most of the work in this field is done when a single outlier is present in the data. However, it is reasonable to suspect more than one outlier, when the number of observations is not too small. Some persons have recommended a sequential procedure for multiple outlier case. But this has the drawback of masking effect, that is, when more than one outlier is present, a test for one outlier may not detect even a single outlier. Hence a block procedure for testing two or more outliers is proposed.

We consider a general linear model, that is $\underset{2}{Y}$ is distributed as normal with mean $\underset{2}{X}\underset{2}{\beta}$ and variance-covariance matrix $\sigma^2\underset{2}{I}$ $(N(\underset{2}{X}\underset{2}{\beta},\sigma^2\underset{2}{I}))$, where $\underset{2}{Y}$ is the n-component vector of random variables, $\underset{2}{\beta}$ is a m-component vector of unknown parameters, σ^2 is the unknown variance of each Y_i, $\underset{2}{X}$ is the known design matrix of order nxm and of rank $k(k \leq m < n)$. The residual vector and the residual sum of squares for this model are $\underset{2}{e} = \underset{2}{\Lambda}\,\underset{2}{y}$ and $S^2 = \underset{2}{y}'\,\underset{2}{\Lambda}\,\underset{2}{y}$, where $\underset{2}{\Lambda} = ((\lambda_{ij})) = [\underset{2}{I} - \underset{2}{X}(\underset{2}{X}'\underset{2}{X})^{-}\,\underset{2}{X}']$ is an idempotent matrix of rank n-k, $\underset{2}{y}$ is the realization of $\underset{2}{Y}$, and $(\underset{2}{X}'\underset{2}{X})^{-}$ is any generalized inverse of $\underset{2}{X}'\underset{2}{X}$.

Suppose s_ν^2 is an independent root mean square estimator of σ^2 based on ν degrees of freedom. Denote the pooled sum of squares based on $p = n-k+\nu$ degrees of freedom by $S_p^2 = S^2 + \nu s_\nu^2$. Define

$$w_i = e_i/[S_p(\lambda_{ii})^{1/2}] \quad , \quad i = 1,2,\ldots,n,$$

as the weighted residuals. We propose the test statistics for two outliers as

$$U = \max_{1\le i<j\le n} u_{ij},$$

and $$V = \max_{1\le i<j\le n} |v_{ij}|,$$

where for $i \neq j$,

$$u_{ij} = (w_i+w_j)/[2(1+\rho_{ij})]^{1/2},$$

$$v_{ij} = (w_i-w_j)/[2(1-\rho_{ij})]^{1/2},$$

and $$\rho_{ij} = \lambda_{ij}/(\lambda_{ii}\,\lambda_{jj})^{1/2}.$$

The joint distribution of these u_{ij}'s as well as v_{ij}'s is obtained. From this the marginal density of u_{ij} is deduced. This is given by

$$f(u_{ij}) = (1-u_{ij}^2)^{(p-3)/2}/B[1/2,(p-1)/2], \quad -1 \le u_{ij} \le 1.$$

Similarly, the joint probability density function (pdf) of u_{ij} and $u_{i_1 j_1}$ is given by

$$(1)\quad f(u_{ij},u_{i_1 j_1}) = \frac{p-2}{2\pi(1-\rho_1^2)^{1/2}}\{1-\frac{1}{1-\rho_1^2}(u_{ij}^2+u_{i_1 j_1}^2 - 2\rho_1\, u_{ij}u_{i_1 j_1})\}^{(p-4)/2}$$

inside the ellipse

$$u_{ij}^2 + u_{i_1 j_1}^2 - 2\rho_1 u_{ij} u_{i_1 j_1} = 1-\rho_1^2 ,$$

where ρ_1 is the shape parameter given by

$$\rho_1 = (\rho_{ii_1}+\rho_{ij_1}+\rho_{i_1 j} + \rho_{jj_1})/[2\{(1+\rho_{ij})(1+\rho_{i_1 j_1})\}^{1/2}].$$

The marginal pdf of v_{ij} is exactly same as that of u_{ij}. Expression for the joint pdf of v_{ij} and $v_{i_1 j_1}$ is analogous to equation (1) with shape parameter ρ_1' given by

$$\rho_1' = (\rho_{ii_1}-\rho_{ij_1} - \rho_{i_1 j} + \rho_{jj_1})/[2\{(1-\rho_{ij})(1-\rho_{i_1 j_1})\}^{1/2}].$$

Since the marginal distributions of u_{ij} and v_{ij} do not depend on i and j, hence the first Bonferroni inequality is useful for evaluating nominal upper percentile points of these statistics. We have also derived a recurrence relation for the evaluation of bivariate probabilities like $\Pr(u_{ij} > h, u_{i_1 j_1} > k)$, as this is required for obtaining bounds for the exact percentile points.

As an application, we consider a random sample from $N(\mu,\sigma^2)$ as a special case of the general regression model. Now $Y_1, Y_2, \ldots, Y_n$ constitute a random sample from a $N(\mu,\sigma^2)$, then for $\nu = 0$, u_{ij} and v_{ij} reduce to

$$u_{ij} = [n/\{2(n-1)(n-2)\}]^{1/2} (Y_i+Y_j - 2\bar{Y})/s$$

and $$v_{ij} = [1/\{2(n-1)\}^{1/2}] (Y_i-Y_j)/s,$$

where $\bar{y} = \sum_{i=1}^{n} y_i/n$, $s^2 = S^2/(n-1)$ and $S^2 = \sum_{i=1}^{n} (y_i-\bar{y})^2$.

This immediately gives

$$U = [n/\{2(n-1)(n-2)\}]^{1/2} T_{N3}$$

and $$V = [1/\{2(n-1)\}^{1/2}] T_{N6},$$

where $T_{N3} = (y_{(n)} + y_{(n-1)} - 2\bar{y})/s$ is the Murphy's statistic and $T_{N6} = (y_{(n)} - y_{(1)})/s$ is the internally studentized range statistic, and $y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(n)}$ are the order statistics obtained from $y_1, y_2, \ldots, y_n$. Both these statistics have been used for the detection of two outliers.

In this case, there are just two values of shape parameter for the joint distribution of u_{ij} and $u_{i_1 j_1}$. These are given by $\rho_1 = (n-4)/[2(n-2)]$ and $\rho_2 = -2/(n-2)$ with respective frequencies $n(n-1)(n-2)/2$ and $n(n-1)(n-2)(n-3)/8$. For the joint distribution of v_{ij} and $v_{i_1 j_1}$, the shape parameter can take three values, viz., -0.5, 0 and 0.5 with respective frequencies $n(n-1)(n-2)/6$, $n(n-1)(n-2)(n-3)/8$ and $n(n-1)(n-2)/3$.

It is shown that the nominal upper percentage points of U and V for some small values of n and ν are exact.

Using the relations between U and T_{N3}, and V and T_{N6}, we obtain nominal percentile points of T_{N3} and T_{N6}. These are compared with the existing tabulated values. These points are in considerable agreement with the tabulated points for

$n \leq 50$. For larger n also, the deviation is not much. A lower bound for the type I error probability is also calculated with the help of the second Bonferroni inequality by evaluating the bivariate probabilities. An approximate percentage point for the Murphy's test statistic T_{N3} is also obtained. By comparing the approximate values with existing tabulated values, we find that the approximation is remarkably good. We, therefore, recommend that the approximate values could be used whenever the exact percentage points are not available, especially for large values of n.

We next consider the case of a two way layout with r rows and c columns and one observation in each cell. Total number of distinct correlation matrices of order 4 for 3x3 and 4x5 tables are enumerated. The total number of distinct shape parameters for 3x3, 4x5 and 5x6 tables along with their respective frequencies are also counted. For the joint distribution of $u_{i_1 j_1, i_2 j_2}$ and $u_{i_3 j_3, i_4 j_4}$ the total number of shape parameters is 40. Their number for 3x3, 4x5 and 5x6 layouts is 11, 37 and 40 respectively, since several of these are equal or are non-existent due to special values of r and c. Similarly, for the joint distribution of $v_{i_1 j_1, i_2 j_2}$ and $v_{i_3 j_3, i_4 j_4}$, the maximum number of distinct shape parameters is 43. Their number for 3x3, 4x5 and 5x6 layouts is 13, 39 and 43 respectively.

These are then used for obtaining bounds for the exact percentile points of U and V statistics. Exact percentage points are also obtained by Monte Carlo method for some combinations of r and c. Nominal percentage points compare favourably with these values.

For studying the performance of the proposed statistics, we assume exactly two outliers are present in the data. The null hypothesis is that there is no outlier and, the alternative hypothesis suitable for the statistic U is the union of $\binom{n}{2}$ hypotheses H_{ij}, $1 \leq i < j \leq n$. Without loss of generality we take H_{12}, where H_{12} states that y_1 and y_2 have a mean shifted to the right by amounts θ_1 and θ_2 respectively. The exact density function of u_{ij} under H_{12} is derived. Our measure of performance is $P_{12} = Pr\,(u_{12} > u_\alpha | H_{12})$, which gives the probability that u_{12} is significantly large under H_{12}. Some other measures are also studied.

The approximate distributions of u_{ij} and v_{ij} under the alternative hypothesis are also obtained. The approximate distribution does not give satisfactory results for evaluating appropriate measures of performance for small values of θ_1 and θ_2 for the statistic U. Consequently, we use exact distribution when θ_1 and θ_2 are small. In comparison to the sequential procedure, our procedure performs better, when there are two outliers in the data.

Extensions for more than two outliers for the statistic U are given. This statistic for $m_1 (\geq 2)$ outliers is given by

$$U = \max_{1 \leq i_1 < i_2 < \ldots < i_{m_1} \leq n} u_{i_1, i_2, \ldots, i_{m_1}},$$

where

$$u_{i_1, i_2, \ldots, i_{m_1}} = (w_{i_1} + w_{i_2} + \ldots + w_{i_{m_1}}) / (m_1 + 2 \sum_{g=1}^{m_1} \sum_{\substack{h=1 \\ i_g < i_h}}^{m_1} \rho_{i_g i_h})^{1/2}.$$

This again reduces to Murphy's statistic for a random sample of size n from a $N(\mu, \sigma^2)$. This statistic $u_{i_1, i_2, \ldots, i_{m_1}}$ has the same univariate density as that of u_{ij} for the two outlier case. For studying the performance of this statistic $\binom{n}{m_1}$ alternative hypotheses have to be considered. The null and the non-null distributions are determined, analogous to that of two outlier case. Performance of this statistic is studied briefly. Such immediate extensions of V do not seem to hold.

CHAPTER I

INTRODUCTION AND SUMMARY

1.1. Scope

The problems of multiple outliers in a fixed effect linear model are considered in this thesis. Outliers are those observations in a sample which deviate in some sense from rest of the observations. Usually these deviations are in mean or in variance or in both. Here we are concerned with test statistics designed to be sensitive to various non-null patterns, primarily a shift in mean of two or more variates when the variance is unknown. The following topics are studied in this thesis:

(i) Detection of two outliers in a general linear model.

(ii) Application to a random sample from $N(\mu, \sigma^2)$.

(iii) Application to a two-way layout.

(iv) Performance of the statistics.

(v) Extensions for more than two outliers.

These topics are discussed in detail with suitable tables to support the theory, wherever necessary. The following sections give an outline of what is covered under these topics.

1.2. Two outliers in a general linear model

Recent work in outlier detection and testing discordancy in a general linear model is based on residuals, standardized

in some way. Anscombe (1961), Anscombe and Tukey (1963) discuss the analysis of residuals and give a detailed presentation of their potential usefulness. Absolute studentized residuals and related statistics for the detection of a single outlier in a general linear model have been considered by several authors, for example, see Srikantan (1961), Stefansky (1971), Joshi (1972, 1975), Ellenberg (1973, 1976), Lund (1975), Prescott (1975) and Gentle (1978). Barnett and Lewis (1978, Ch. 7), Kale (1979), Hawkins (1980, Ch. 7), David (1981, Section 8.6) and Beckman and Cook (1983) have done excellent survey work in this field. More recent work is done by Cook and Prescott (1981) and Doornbos (1980, 1981). All these authors have mainly concentrated on a single outlier, with some also considering sequential and other precedures for two or more outliers. Their work is mainly based on the assumption that there is at most one outlier in the given data set, an assumption which is reasonable when the number of observations is small. However, it is reasonable to suspect more than one outlying observations, when the number of observation is not too small. Some authors, for example, Anscombe (1960), John and Draper (1978), Gentleman (1980) etc. have recommended a sequential approach for detecting more than one outlier. However, in the special case of a random sample from a $N(\mu,\sigma^2)$ distribution, with two outliers on the right, McMillan (1971) and Moran and McMillan (1973) have shown that the performance of such tests is inferior to that of Murphy's test (Murphy,1951) for two outliers.

In general the total number of outliers present in any given data is unknown. However, there are situations where one feels that a specified number of observations are outliers, for example, there may be a sudden drop in output of two machines out of n machines in a cloth manufacturing factory. If there are two outlying observations, then a statistic which can detect both the outliers simultaneously is preferred over a sequential procedure, since it avoids the masking effect. By masking, it is meant that "extreme" observations are not declared as outliers because some other observations are also outliers. Thus if there are two outliers in a data set and we test for one outlier, then we may not be able to detect any outlier due to the presence of second outlier. This phenomenon is discussed by Pearson and Chandra Sekar (1936), McMillan and David (1971), McMillan (1971) etc. With similar motivation, we consider block procedures for testing two outliers in linear models.

Let $Y_1, Y_2, \ldots, Y_n$ be n independently and normally distributed random variables which have a linear regression on a known set of m variables. Then the fixed effects linear model can be described as

$$\underset{\sim}{Y} \overset{d}{=} N(X\underset{\sim}{\beta}, \sigma^2 I), \tag{1.2.1}$$

where the symbol "$\overset{d}{=}$" stands for "is distributed according to",

$$\underset{2}{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \quad ; \quad \underset{2}{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix} \quad ; \quad \underset{2}{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & & & \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix};$$

$\underset{2}{I}$ is the identity matrix of order n, and $\underset{2}{\beta}$ and σ^2 are the unknown parameters. We also assume that $m < n$, that is, there are more observations than the number of unknown parameters in $\underset{2}{\beta}$.

Let the rank of the known design matrix $\underset{2}{X}$ be $k \leq m$, and $\underset{2}{y}$ stand for the realization of $\underset{2}{Y}$. Then the normal equations for this model are givenby

$$\underset{2}{X}' \underset{2}{X} \underset{2}{\beta} = \underset{2}{X}' \underset{2}{y}.$$

Let $(\underset{2}{X}'\underset{2}{X})^-$ denote any generalized inverse of $(\underset{2}{X}'\underset{2}{X})$ satisfying $(\underset{2}{X}'\underset{2}{X})(\underset{2}{X}'\underset{2}{X})^-(\underset{2}{X}'\underset{2}{X}) = \underset{2}{X}'\underset{2}{X}$, Rao (1973, p. 24). Then one solution of the normal equations is

$$\hat{\underset{2}{\beta}} = (\underset{2}{X}'\underset{2}{X})^- \underset{2}{X}'\underset{2}{y}.$$

The estimated value of $\underset{2}{y}$ for this model is

$$\hat{\underset{2}{y}} = \underset{2}{X}\hat{\underset{2}{\beta}} = \underset{2}{X}(\underset{2}{X}'\underset{2}{X})^- \underset{2}{X}'\underset{2}{y}$$

and the residual vector is

$$\underset{2}{e} = \underset{2}{y} - \hat{\underset{2}{y}} = [\underset{2}{I} - \underset{2}{X}(\underset{2}{X}'\underset{2}{X})^-\underset{2}{X}']\underset{2}{y} = \underset{2}{\Lambda}\underset{2}{y}, \tag{1.2.2}$$

where

$$\underset{2}{\Lambda} = \underset{2}{I} - \underset{2}{X}(\underset{2}{X}'\underset{2}{X})^- \underset{2}{X}' = ((\lambda_{ij})) \tag{1.2.3}$$

is a real, symmetric and idempotent matrix of rank (n-k) satisfying $\underset{\sim}{\Lambda}\,\underset{\sim}{X} = \underset{\sim}{O}$. The residual vector $\underset{\sim}{e}$ has a singular normal distribution $N(\underset{\sim}{O}, \underset{\sim}{\Lambda}\,\sigma^2)$. Further, the residual sum of square

$$S^2 = \underset{\sim}{e}'\underset{\sim}{e} = \underset{\sim}{y}'\,\underset{\sim}{\Lambda}\,\underset{\sim}{y} \qquad (1.2.4)$$

has a central $\sigma^2\chi^2$ distribution with (n-k) degrees of freedom.

Let s_ν be a root mean square estimator of σ based on ν degrees of freedom, which is independent of $\underset{\sim}{y}$ and $S_p^2 = S^2 + \nu s_\nu^2$ be the pooled sum of squares based on $p = n-k+\nu$ degrees of freedom. Define

$$w_i = e_i/(S_p\,\lambda_{ii}^{1/2}), \quad i = 1,2,\ldots,n \qquad (1.2.5)$$

as weighted residuals. The joint destribution of these residuals has been discussed by Joshi (1972, 1975). This is generalized by Ellenberg (1973) for the non-singular joint distribution of $(w_1,\ldots,w_s)$. An ingenious method for obtaining these distributions is also given by Margolin (1977).

In Chapter II, using linear combinations of these residuals, we suggest some test statistics for detection of two outliers. We denote the one-sided statistic by U and the two-sided statistic by V. These are given by

$$U = \max_{1\le i<j\le n} u_{ij}$$

and

$$V = \max_{1\le i<j\le n} |v_{ij}|$$

where for $i \neq j$, we assume $|\rho_{ij}| < 1$, and the maximum occu
for a single pair (i,j), and

(1.2.6) $$u_{ij} = (w_i+w_j)/[2(1+\rho_{ij})]^{1/2},$$

(1.2.7) $$v_{ij} = (w_i-w_j)/[2(1-\rho_{ij})]^{1/2}, \text{ and}$$

$$\rho_{ij} = \lambda_{ij}/(\lambda_{ii}\lambda_{jj})^{1/2}$$

is the correlation coefficient between the residuals e_i and e_j.

The one-sided statistic U can be used for two outliers on the right side. For two outliers on left, a suitable test statistic is

$$U_1 = \min_{1\le i<j\le n} u_{ij}.$$

The distribution properties of U_1 are analogous to that of U. Hence it is sufficient to study U for two outliers on the right. In order to apply these tests we require the exact null distribution of these statistics when there are no outliers. This is extremely complicated and among other things, depends on the design matrix $\underset{\sim}{X}$. However, the marginal distribution of u_{ij} and v_{ij} are identical. The common distribution is given by

$$f(u_{12}) = (1-u_{12}^2)^{(p-3)/2}/B[1/2,(p-1)/2] \ , \ -1 \le u_{12} \le 1.$$

Using first Bonferroni inequality this allows us to obtain an upper bound for the true upper percentage point. This gives nominal upper percentage point, which controls the probability of type I error. A lower bound is obtained by considering the joint distribution of u_{ij}'s in some special cases. This requires the evaluation of bivariate probabilities like $Pr(u_{ij}>c, u_{i'j'}>c)$. In this regard we have provided some recurrence relations for

evaluation of such bivariate probability terms.

A two sided discordancy test for m_1 outliers (irrespective of directions) in a random sample from $N(\mu,\sigma^2)$ is studied by Tietjen and Moore (1972). Their statistic called the "largest gap", which is denoted as T_{N16} by Barnett and Lewis (1978), is given by

$$(1.2.8) \qquad T_{N16} = \sum_{j=1}^{n-m_1} (r_{(j)} - \bar{r}_{n-m_1})^2 / \sum_{j=1}^{n} (r_{(j)} - \bar{r})^2 .$$

Here $r_{(j)}=|y_j-\bar{y}|$, the absolute deviation of y_j from the sample mean $\bar{y}$; $\{r_{(j)}\}$ are the values of r_j in ascending order, $r_{(1)} < r_{(2)} < \ldots < r_{(n)}$; $\bar{r}$ is the mean of all the r_j's; and

$$\bar{r}_{n-m_1} = (r_{(1)} + r_{(2)} + \ldots + r_{(n-m_1)})/(n-m_1).$$

For the general linear model considered above, an obvious generalization using weighted residuals is as follows. Let

$r_i^* = |w_i|$, $r_{(1)}^* < r_{(2)}^* < \ldots < r_{(n)}^*$ be ordered r_i^*'s ,

$\bar{r}^* = \sum_{i=1}^{n} r_i^*/n$ and $\bar{r}_{n-m_1}^* = (r_{(1)}^* + \ldots + r_{(n-m_1)}^*)/(n-m_1)$.

Then

$$T_{N16}^* = \sum_{j=1}^{n-m_1} (r_{(j)}^* - \bar{r}_{n-m_1}^*)^2 / \sum_{j=1}^{n} (r_{(j)}^* - \bar{r}^*)^2$$

can be used for detection of m_1 outliers when the direction is unknown. Note that for a random sample of size n from $N(\mu,\sigma^2)$, we have $e_i = y_i - \bar{y}$ and

$$\lambda_{ii} = \lambda = (n-1)/n \quad \text{for all } i = 1,\ldots,n.$$

Consequently

$$r^{*}_{(j)} = r_{(j)}/(S_p \lambda^{1/2}),\ \bar{r}^{*}_{n-m_1} = \bar{r}_{n-m_1}/(S_p \lambda^{1/2}),\ \text{and}$$

$$T^{*}_{N16} = \frac{\sum_{j=1}^{n-m_1} (r_{(j)} - \bar{r}_{n-m_1})^2/(S_p^2 \lambda)}{\sum_{j=1}^{n} (r_{(j)} - \bar{r})^2/(S_p^2 \lambda)}$$

$$= T_{N16}.$$

Because of the modulus sign involved in this statistic, it is not easily comprehensible and is very complicated to deal with. Also as pointed out by David (1981, p. 240), even T_{N16} does not necessarily pick up the m_1 most outlying observations.

Shapiro and Wilk (1965) have also proposed a statistic for random samples from $N(\mu,\sigma^2)$, which we denote again in Barnett and Lewis's (1978) notation as T_{N17}. This is given by

$$T_{N17} = \Big[\sum_{i=1}^{[n/2]} a_{n,n-i+1} \{y_{(n-i+1)} - y_{(i)}\}\Big]^2/S^2,$$

where $[n/2]$ denotes the integer part of $n/2$. $a_{n,j}$ are tabulated constants, and $y_{(1)} \le y_{(2)} \le \cdots \le y_{(n)}$ are ordered observations $y_1, y_2, \ldots, y_n$.

For linear models one can use a generalized statistic

$$T^{*}_{N17} = \sum_{i=1}^{[n/2]} a^{*}_{n,n-i+1} \left[w_{(n-i+1)} - w_{(i)} \right]^2 ,$$

where $w_{(1)} \le w_{(2)} \le \cdots \le w_{(n)}$ are the ordered $w_1, w_2, \ldots, w_n$ and $a^{*}_{n,n-i+1}$ are suitable constants.

These statistics T_{N16} and T_{N17} have not been studied in complete detail so far. The generalizations T^{*}_{N16} and T^{*}_{N17} will require extensive calculations even for well behaved patterned design matrices $\underset{\sim}{X}$. For this reason, we have not considered such statistics and have concentrated on U and V statistics only.

1.3. Application to a random sample from a $N(\mu,\sigma^2)$

In Chapter III we consider a random sample $y_1, y_2, \ldots, y_n$ of size n from a normal population with mean μ and variance σ^2. Let

$$y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(n)}$$

be the order statistics, obtained by arranging $y_1, y_2, \ldots, y_n$ in an ascending order of magnitude.

This can be considered as a special case of the general regression model and the test statistics U and V can be obtained from the regression model case. For $\nu = 0$, the one-sided statistic U can be compared with the Murphy's statistic M for two outliers, which is given by

$$(1.3.1) \qquad M = (y_{(n)} + y_{(n-1)} - 2\bar{y})/S$$

where

$$(1.3.2) \qquad S^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 .$$

The two-sided statistic V can be compared with internally studentized range

$$(1.3.3) \qquad T_{N6} = (y_{(n)} - y_{(1)})/s ,$$

where $s^2 = S^2/(n-1)$.

We show that for $\nu = 0$, our statistics U and V, when multiplied with suitable constants reduce to M and T_{N6} respectively. Using nominal percentile points of U and V, nominal critical values of M and T_{N6} are calculated. These nominal critical values of M are compared with the exact values tabulated by Hawkins (1978) for Murphy's test. The T_{N6} values are compared with the simulated values obtained by Barnett and Lewis (1978). It is observed that the nominal percentage points are reasonable for $n \leq 50$.

We also obtain a lower bound for type I error probability using second Bonferroni inequality. These are tabulated for different values of n and significance level α.

Finally, a method for finding approximate upper percentiles of Murphy's test statistic for two outliers is discussed. Comparing with the tabulated values, it is observed that the approximation is remarkably good for all values of n.

1.4. Application to a two-way layout

Tests of multiple outliers in a two way-layout have been discussed by several authors like Gentleman and Wilk (1975a,b), John and Draper (1978), Bradu and Hawkins (1982) etc. Some of them have used residuals while others have used tetrads etc. In Chapter IV we apply statistics U and V for the detection of two outliers in a two-way table having a single observation in each cell.

In this chapter we mainly analyse the 'shape parameters' which appear in the joint distribution of two u_{ij}'s or two v_{ij}'s. These shape parameters are then used for finding bounds for actual percentile points. For comparison purposes, actual percentile points are also obtained by Monte Carlo method for some special cases.

1.5. Performance of the statistics

In Chapter V we study the performance of test statistics in non-null situation when two outliers are present. We now assume that exactly two outliers are present. The null hypothesis for testing for two outliers specifies that there are no outliers.

To evaluate the performance of the statistic U the alternative hypothesis is the union of $\binom{n}{2}$ hypotheses H_{ij} $(1 \leq i < j \leq n)$. Without loss of generality we take H_{12}. For testing two outliers on right, the model under H_{12} is

$$E(\underset{\sim}{y}) = \underset{\sim}{X}\,\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}_1\,\theta_1 + \underset{\sim}{\varepsilon}_2\,\theta_2 ,$$

where $\underset{\sim}{\varepsilon}_s$ $(s = 1,2,\ldots,n)$ is the sth column of $\underset{\sim}{I}_n$ and $\theta_i > 0$, $i = 1,2$.

For determing the performance of the two-sided statistic V, the model under H_{12}^{*} is given by

$$E(\underset{\sim}{y}) = \underset{\sim}{X}\,\underset{\sim}{\beta} - \underset{\sim}{\varepsilon}_1\,\theta_1 + \underset{\sim}{\varepsilon}_2\,\theta_2 ;\ \theta_1,\theta_2 > 0.$$

In this case the alternative hypothesis is the union of $n(n-1)$ such hypotheses.

We have obtained the exact and approximate distribution of u_{ij}'s and v_{ij}'s under the alternative hypothesis. For U, the approximate method does not give satisfactory results for small values of θ_1 and θ_2. Consequently, its performance is evaluated with the exact density function for small θ_1 and θ_2. For large θ_1, θ_2 we can evaluate it approximately also. Let

$$p^{i,j} = \Pr\,(u_{ij} > u_\alpha | H_{12}), \text{ and}$$

$$p^{*i,j} = \Pr(|v_{ij}| > v_\alpha | H_{12}^*).$$

We use these and related quantities for studying the performance of proposed test statistics. These measures are calculated for random samples from $N(\mu,\sigma^2)$ for several values of n.

Similarly this is carried over for two way tables also. We first determine the cells which would give a minimum probability of identification of outliers, if they are present in them. Then we evaluate the measures used for studying the performance.

For comparison purposes we consider the sequential procedure. The possibility of testing multiple outliers sequentially for discordancy has been mentioned by a number of authors like Pearson and Chandra Sekar (1936), Dixon (1953), Tietjen and Moore (1972) and David (1981, p. 239). The sequential procedure with other block procedures has been compared by McMillan (1971), McMillan and David (1971), and Moran and McMillan (1973).

We compare our procedure in case of a random sample from normal distribution with that of sequential procedure whose exact performance values are obtained by Moran and McMillan (1973). For two way tables we consider the sequential procedure suggested by Anscombe (1960). For all these cases our procedure performs better than the sequential procedure.

1.6. Extension for more than two-outlier cases.

The one-sided statistic U has been extended for more than two outliers, which from a random sample for $N(\mu,\sigma^2)$ is equivalent to Murphy's test statistic. Distributions in the null case are analogous to that of two-outlier case. We have tabulated the nominal percentile points for three-outlier case.

For studying the performance of the statistic for 3 outliers on right, we have $\binom{n}{3}$ alternative hypotheses, H_{hij} $(1 \leq h < i < j \leq n)$, where under H_{123}

$$E(\underset{\sim}{y}) = \underset{\sim}{X}\,\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}_1\,\theta_1 + \underset{\sim}{\varepsilon}_2\,\theta_2 + \underset{\sim}{\varepsilon}_3\,\theta_3 ,$$

$\underset{\sim}{\varepsilon}_s$ $(s = 1,2,\ldots,n)$ is the sth column of $\underset{\sim}{I}_n$, $\theta_i > 0$ $(i = 1,2,3)$ and

$$\text{Var}\,(\underset{\sim}{y} \mid H_{123}) = \sigma^2 \underset{\sim}{I}.$$

Distribution of test statistic in the non-null case is discussed in detail. Again the results are analogous to the two-outlier case. We have studied the performance of Murphy's test statistic for three-outlier case. It is observed that the test for three outliers performs well when outliers of the same

magnitude are present. If there are only two outliers, and a test for three outliers is applied, then the performance is not good for small values of n. But it is reasonable for large values of n. However, if there is only one outlier and a test for three outliers is applied, then it performs very poorly even for large values of n.

General results for m_1 (> 3) outliers are similar. We also note that such extension for three or more outliers are possible for the one-sided statistic U only. No immediate extension seems to hold for the two-sided case.

1.7. Notations

The following notations will be used in this thesis :

pdf : probability density function,

$\overset{d}{=}$: is distributed according to,

$\cong$: is approximately equal to,

$\equiv$: is equivalent to,

$G(a) = \int_0^\infty x^{a-1} e^{-x} dx,\ a > 0,$

$B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1} dx,\ a > 0,\ b > 0,$

$I_y(a,b) = \int_0^y x^{a-1} (1-x)^{b-1} dx/B(a,b)$,

$N(\mu,\sigma^2)$: normal distribution with mean μ and variance σ^2,

χ^2_a : central chi-square distribution with 'a' degrees of freedom

and

$\chi^2(a,b)$: non-central chi-square distribution with 'a' degrees of freedom and non-centrality parameter 'b'.

CHAPTER II

TWO OUTLIERS IN A GENERAL LINEAR REGRESSION MODEL

2.1. Introduction and test statistics

The problem of detecting two outliers in a general linear model is considered in this chapter. Let $Y_1, Y_2, \ldots, Y_n$ be n independently and normally distributed random variables and $y_1, y_2, \ldots, y_n$ be the realization of these variables. We consider the model as described in equation (1.2.1) of Section 1.2. Then we have the residual vector $\underset{\sim}{e}$, the variance-covariance matrix $\underset{\sim}{\Delta}$ of $\underset{\sim}{e}$ and the residual sum of squares S^2 as in equations (1.2.2), (1.2.3) and (1.2.4) respectively. The random variables used for the detection of a single outlier are the weighted residuals w_i given at equation (1.2.5), viz.

$$w_i = e_i/(S_p \lambda_{ii}^{1/2}), \quad i = 1,2,\ldots,n,$$

where $S_p^2 = S^2 + \nu s_\nu^2$ is a pooled sum of squares based on $p = n-k+\nu$ degrees of freedom.

The statistics $\underset{1\le i\le n}{\text{Max}} \; w_i$ and $\underset{1\le i\le n}{\text{Max}} \; |w_i|$ are used for the detection of a single outlier, for example, see Srikantan (1961), Joshi (1972, 1975), Ellenberg (1973,1976), Lund (1975), Cook and Prescott (1981), and Doornbos (1981). For the case of two outliers, we propose a linear combination of w_i and w_j. The proposed statistics generalize the well known Murphy's test and

a test based on studentized range for the case of a random sample from $N(\mu,\sigma^2)$ population. For $i \neq j = 1,2,\ldots,n$, define u_{ij} and v_{ij} by

$$(2.1.1) \quad u_{ij} = (w_i + w_j)/[2(1+\rho_{ij})]^{1/2} \text{ and}$$

$$v_{ij} = (w_i - w_j)/[2(1-\rho_{ij})]^{1/2}$$

where ρ_{ij} is the correlation coefficient between e_i and e_j and is given by

$$(2.1.2) \quad \rho_{ij} = \lambda_{ij}/(\lambda_{ii}\,\lambda_{jj})^{1/2}.$$

We now propose the test statistics for two outliers as

$$U = \max_{1\leq i<j\leq n} u_{ij} \quad \text{and} \quad V = \max_{1\leq i<j\leq n} |v_{ij}|.$$

The statistic U is useful for detecting two outliers on right, while V is useful for detecting two outliers on either direction.

For two outliers on left, a suitable test statistic is given by

$$U_1 = -\min_{1\leq i<j\leq n} u_{ij}.$$

The distribution properties of U_1 are analogous to that of U. Hence we consider only U in detail.

Thus to detect two outliers, we compute all the u_{ij}'s and v_{ij}'s and find the corresponding U, V and U_1. If the statistic U exceeds the critical value u_α at α level of significance, then the two observations corresponding to the maximum u_{ij} are

declared as outliers on the right side; if we suspect that the two outliers are present on either end, then the maximum of absolute values of v_{ij} would identify them; and for outliers on left the observations corresponding to the minimum of u_{ij}'s would identify them.

Example 2.1.1. As an illustration of our procedure, consider the hypothetical data given in Table 2.1.1 of a 4x5 layout given by Daniel (1960) and discussed by Bross (1961), Mickey et al. (1967) and Doornbos (1981). Let y_{ij} ($i = 1,2,\ldots,4$ and $j = 1,2,\ldots,5$) denote the observations of this two-way table. Residual for (i,j)th cell will be denoted by e_{ij} ($i = 1,2,\ldots,4$ $j = 1,2,\ldots,5$), and also by single subscripts, viz.

$$\underset{\sim}{e}' = (e_1, e_2, \ldots, e_{20}) = (e_{11}, \ldots, e_{15}, e_{21}, \ldots, e_{25}, \ldots, e_{41}, \ldots, e_{45}).$$

TABLE 2.1.1. Hypothetical yields.

Levels of A	Levels of B				
	1	2	3	4	5
1	35	29	25	19	22
2	32	29	29	25	20
3	37	34	30	25	29
4	40	36	20	35	29

The corresponding residuals are given by

2	0	2	-4	0
-2	-1	5	1	-3
-1	0	2	-3	2
1	1	-9	6	1

For this data $S^2 = 202.0$. Further, the variance of each residual is $\lambda\sigma^2$ where $\lambda = 12/20$. The two largest positive residuals are e_{23} and e_{44}, while others are much smaller. Consequently, these two are likely to give the maximum value of u_{ij}. It is indeed so, and the value of U (for $\nu = 0$) is given by

$$U = \max_{1\le i<j\le n} u_{ij} = [1/(S_p \lambda^{1/2})] \max_{1\le i<j\le n} (e_i+e_j)/(2+2\rho_{ij})^{1/2}$$

$$= (5+6)/[(202\times 0.6)^{1/2} (2+2/12)^{1/2}]$$

$$= 0.6788.$$

The observations which give the maximum of u_{ij}'s are y_{23} and y_{44}. Similarly, V is given by

$$V = \max_{1\le i<j\le n} |v_{ij}| = [1/(S_p \lambda^{1/2})] \max_{1\le i<j\le n} |(e_i-e_j)/(2-2\rho_{ij})^{1/2}|$$

$$= (6+9)/[(202\times 0.6)^{1/2} (2+2/4)^{1/2}]$$

$$= 0.8617.$$

The corresponding observations which give rise to V are y_{43} and y_{44}.

Similarly, the value of U_1 is

$$U_1 = -\min_{1\le i<j\le n} u_{ij} = -[1/(S_p \lambda^{1/2})] \min_{1\le i<j\le n} (e_i+e_j)/(2+2\rho_{ij})^{1/2}$$

$$= -(-9-4)/[(202\times 0.6)^{1/2}(2+2/12)^{1/2}]$$

$$= 0.8022.$$

The observations corresponding to minimum of u_{ij}'s are y_{14}

and y_{43}. Depending upon the model under consideration, these observations are prime suspected outliers. Thus y_{23} and y_{44} on right side, y_{43} and y_{44} on either side and y_{14} and y_{43} on left side are the potential outliers.

2.2. Distribution theory

We will first obtain the marginal densities of u_{ij}'s and v_{ij}'s and then consider the joint densities. Without loss of generality, we derive the marginal pdf of u_{12} and v_{12}. The bivariate density of w_1 and w_2 as defined in (1.2.5) is given by Joshi (1972).

$$(2.2.1)\quad g(w_1, w_2) = \frac{p-2}{2\pi(1-\rho^2)^{1/2}} \left(1 - \frac{w_1^2 - 2\rho w_1 w_2 + w_2^2}{1-\rho^2}\right)^{(p-4)/2},$$

where $\rho = \rho_{12}$ and the region of positive density is the interior of the ellipse

$$w_1^2 - 2\rho w_1 w_2 + w_2^2 = 1-\rho^2.$$

For finding the joint probability density function (pdf) of u_{12} and v_{12}, consider the transformation

$$u = (w_1+w_2)/[2(1+\rho)]^{1/2} = a\,(w_1+w_2)$$

and $$v = (w_1-w_2)/[2(1-\rho)]^{1/2} = b(w_1-w_2),$$

where $a = 1/[2(1+\rho)]^{1/2}$ and $b = 1/[2(1-\rho)]^{1/2}$.

The inverse transformation is

$$w_1 = (u/a + v/b)/2,$$

$$w_2 = (u/a - v/b)/2$$

and the jacobian of transformation in absolute value is given by

$$|J| = \left|\frac{\partial(w_1, w_2)}{\partial(u,v)}\right| = \begin{vmatrix} 1/(2a) & 1/(2b) \\ 1/(2a) & -1/(2b) \end{vmatrix} = 1/(2ab) \ .$$

Consequently, the joint pdf of $u_{12} = u$ and $v_{12} = v$ is

$$g(u,v) = \frac{p-2}{4\pi(1-\rho^2)^{1/2}\,ab}\left[1 - \frac{1}{1-\rho^2}\left\{\frac{1}{4}\left(\frac{u}{a}+\frac{v}{b}\right)^2 - 2\cdot\frac{1}{4}\left(\frac{u}{a}+\frac{v}{b}\right)\left(\frac{u}{a}-\frac{v}{b}\right)\rho + \frac{1}{4}\left(\frac{u}{a}-\frac{v}{b}\right)^2\right\}\right]^{(p-4)/2} .$$

Substituting for a and b and simplifying we get

$$(2.2.2) \qquad g(u,v) = \frac{p-2}{2\pi}(1-u^2-v^2)^{(p-4)/2}, \ u^2+v^2 \leq 1.$$

Integrating out v, we get the marginal pdf of u as

$$g(u) = \frac{p-2}{2\pi}\int_{-(1-u^2)^{1/2}}^{(1-u^2)^{1/2}} (1-u^2-v^2)^{(p-4)/2}\, dv.$$

Now substituting $t = v/(1-u^2)^{1/2}$, we have

$$g(u) = \frac{p-2}{2\pi}(1-u^2)^{(p-3)/2}\int_{-1}^{1}(1-t^2)^{(p-4)/2}\, dt$$

$$(2.2.3) \qquad = (1-u^2)^{(p-3)/2}/B[1/2,(p-1)/2] \ , \ -1 \leq u \leq 1.$$

Due to symmetry in equation (2.2.2) with respect to u and v the marginal pdf of v_{12} is exactly same as that of u_{12}. It should be noted that the marginal pdf of w_i is also same as that of u_{12} given at equation (2.2.3).

For finding the joint pdf of two or more u_{ij}'s or v_{ij}'s, it is possible to proceed on analogous lines. We start with the

joint distribution of $(w_1, w_2, \ldots, w_s)$ as given by Ellenberg (1973), and obtain the desired joint distribution by means of suitable transformations. However, the integration of extra variables becomes tedious. We therefore proceed on lines similar to Ellenberg, using independence of certain quadratic forms. The main result is given in Theorem 2.2.1. We need the following lemma.

Lemma 2.2.1.

Let the residual vector $\underset{\sim}{e}$ and the residual sum of squares S^2 be as given by equation (1.2.2) and (1.2.4) respectively. Define

$$\underset{\sim}{z} = (e_1/\lambda_{11}^{1/2}, \ldots, e_n/\lambda_{nn}^{1/2})' = \underset{\sim}{D}\,\underset{\sim}{e},$$

where $\underset{\sim}{D}$ is a diagonal matrix given by

$$\underset{\sim}{D} = \text{diag}\,(\lambda_{11}^{-1/2}, \lambda_{22}^{-1/2}, \ldots, \lambda_{nn}^{-1/2}). \quad \text{Let}$$

$$(2.2.4) \qquad \underset{\sim}{D}\,\underset{\sim}{\Lambda}\,\underset{\sim}{D}' = \underset{\sim}{R} = ((\rho_{ij}))$$

be the correlation matrix of the residual vector $\underset{\sim}{e}$. Let $\underset{\sim}{M}$ be a sxn matrix such that $\underset{\sim}{C} = \underset{\sim}{M}\,\underset{\sim}{R}\,\underset{\sim}{M}'$ is positive definite. Consider a linear transformation

$$\underset{s\times 1}{\underset{\sim}{T}} = \underset{s\times n}{\underset{\sim}{M}}\ \underset{n\times 1}{\underset{\sim}{z}}\ .$$

Then

(i) $\underset{\sim}{T}$ is distributed as s-variate $N(\underset{\sim}{0}, \underset{\sim}{C}\,\sigma^2)$ variate.

(ii) S_1^2/σ^2 is distributed as a χ^2 variate with (n-k-s) degrees of freedom, where

$$S_1^2 = S^2 - T_2' C_2^{-1} T_2$$

(iii) S_1^2 and T_2 are independently distributed.

Proof : Since $e_2 \overset{d}{=} N(O, \Lambda_2 \sigma^2)$, hence

$$z_2 = D_2 e_2 \overset{d}{=} N(O, D_2 \Lambda_2 D_2' \sigma^2), \text{ that is } N(O, R_2 \sigma^2).$$

Consequently,

$$\underset{s\times 1}{T_2} = \underset{s\times n}{M_2} \quad \underset{n\times 1}{z_2} \overset{d}{=} N_s(O, C_2 \sigma^2) .$$

This distribution is non-singular due to the assumption that C_2 is positive definite.

Next consider the quadratic form

$$\begin{aligned} T_2' C_2^{-1} T_2 &= z_2' M_2' C_2^{-1} M_2 z_2 \\ &= e_2' D_2' M_2' C_2^{-1} M_2 D_2 e_2 \\ &= y_2' \Lambda_2' D_2' M_2' C_2^{-1} M_2 D_2 \Lambda_2 y_2 \\ &= y_2' B_2 y_2 , \text{ (say)}, \end{aligned}$$

where

$$B_2 = \Lambda_2' D_2' M_2' C_2^{-1} M_2 D_2 \Lambda_2 \tag{2.2.5}$$

is a real symmetric matrix with

$$B_2 \Lambda_2 = \Lambda_2' D_2' M_2' C_2^{-1} M_2 D_2 \Lambda_2 \Lambda_2 = B_2.$$

Similarly, $\Lambda_2 B_2 = B_2$. Further

$$\begin{aligned} B_2 B_2 &= \Lambda_2 D_2' M_2' C_2^{-1} M_2 D_2 \Lambda_2 \Lambda_2 D_2' M_2' C_2^{-1} M_2 D_2 \Lambda_2 \\ &= \Lambda_2 D_2' M_2' C_2^{-1} C_2 C_2^{-1} M_2 D_2 \Lambda_2 = B_2, \end{aligned}$$

on using that Λ_2 is idempotent and $D_2 \Lambda_2 D_2' = R_2$.

Thus B is an idempotent matrix with rank given by

$$\text{rank}(B) = \text{tr}(B) = \text{tr}(\Lambda D' M' C^{-1} M D \Lambda)$$
$$= \text{tr}(M D \Lambda \Lambda D' M' C^{-1}) = \text{tr}(C C^{-1})$$
$$= \text{tr}(I_s) = s.$$

This at once gives that

$$\frac{1}{\sigma^2} y' B y = \frac{1}{\sigma^2} T' C^{-1} T \overset{d}{=} \chi^2_s \text{ (central)}$$

since the noncentrality parameter is

$$\frac{1}{\sigma^2} \beta' X' B X \beta = \frac{1}{\sigma^2} \beta' X' B \Lambda X \beta = 0, \text{ as}$$

$B = B \Lambda$ and $\Lambda X = 0$.

The distribution of $T' C^{-1} T$ can be obtained directly as well, for example, see Rao (1973, p. 524).

Next, consider

$$S_1^2 = S^2 - T' C^{-1} T$$
$$= S^2 - y' B y = y' \Lambda y - y' B y$$
$$= y'(\Lambda - B) y.$$

The matrix $(\Lambda - B)$ is an idempotent matrix, since $\Lambda B = B \Lambda = B$ and Λ and B are idempotent matrices. Further, rank of $(\Lambda - B)$ is $(n-k-s)$.

Thus $S_1^2/\sigma^2 \overset{d}{=} \chi^2_{n-k-s}$,

where the non-centrality parameter is again zero.

Now, consider the decomposition

$$\underset{\sim}{y}'\underset{\sim}{y} = \underset{\sim}{y}'(\underset{\sim}{I}-\underset{\sim}{\Lambda})\underset{\sim}{y} + \underset{\sim}{y}'(\underset{\sim}{\Lambda}-\underset{\sim}{B})\underset{\sim}{y} + \underset{\sim}{y}'\ \underset{\sim}{B}\ \underset{\sim}{y},$$

since $n = \text{rank}\ (\underset{\sim}{I} - \underset{\sim}{\Lambda}) + \text{rank}\ (\underset{\sim}{\Lambda} - \underset{\sim}{B}) + \text{rank}\ (\underset{\sim}{B})$, hence by Fisher-Cochran Theorem it follows that the quadratic forms appearing on the R.H.S. are independent. In particular $S_1^2 = \underset{\sim}{y}'(\underset{\sim}{\Lambda} - \underset{\sim}{B})\underset{\sim}{y}$ and $\underset{\sim}{y}'\ \underset{\sim}{B}\ \underset{\sim}{y}$ are independent. This is even otherwise obvious, since $(\underset{\sim}{\Lambda} - \underset{\sim}{B})\ \underset{\sim}{B} = \underset{\sim}{\Lambda}\underset{\sim}{B} - \underset{\sim}{B}^2 = \underset{\sim}{B} - \underset{\sim}{B} = \underset{\sim}{O}$.

The independence of $S_1^2 = \underset{\sim}{y}'(\underset{\sim}{\Lambda} - \underset{\sim}{B})\underset{\sim}{y}$, and $\underset{\sim}{T} = \underset{\sim}{M}\ \underset{\sim}{D}\ \underset{\sim}{\Lambda}\ \underset{\sim}{y}$ is also immediate, since

$$\begin{aligned}
\underset{\sim}{M}\ \underset{\sim}{D}\ \underset{\sim}{\Lambda}\ (\underset{\sim}{\Lambda} - \underset{\sim}{B}) &= \underset{\sim}{M}\ \underset{\sim}{D}\ (\underset{\sim}{\Lambda} - \underset{\sim}{\Lambda}\ \underset{\sim}{B}) = \underset{\sim}{M}\ \underset{\sim}{D}\ \underset{\sim}{\Lambda} - \underset{\sim}{M}\ \underset{\sim}{D}\ \underset{\sim}{B} \\
&= \underset{\sim}{M}\ \underset{\sim}{D}\ \underset{\sim}{\Lambda} - \underset{\sim}{M}\ \underset{\sim}{D}\ \underset{\sim}{\Lambda}'\ \underset{\sim}{D}'\ \underset{\sim}{M}'\ \underset{\sim}{C}^{-1}\ \underset{\sim}{M}\ \underset{\sim}{D}\ \underset{\sim}{\Lambda} \\
&= \underset{\sim}{M}\ \underset{\sim}{D}\ \underset{\sim}{\Lambda} - \underset{\sim}{C}\ \underset{\sim}{C}^{-1}\ \underset{\sim}{M}\ \underset{\sim}{D}\ \underset{\sim}{\Lambda} \\
&= \underset{\sim}{M}\ \underset{\sim}{D}\ \underset{\sim}{\Lambda} - \underset{\sim}{M}\ \underset{\sim}{D}\ \underset{\sim}{\Lambda} = \underset{\sim}{O}.
\end{aligned}$$

This completes the proof of the lemma.

The notations used in Lemma 2.2.1 are also used in the following theorem, which gives the desired joint pdf.

<u>Theorem 2.2.1</u>. Let $S_p^2 = S^2 + \nu s_\nu^2$, $p = n-k+\nu$, where s_ν is an independent root mean square estimator of σ. Let

$$u_i = t_i/S_p,\ i = 1,2,\ldots,s,$$

where

$$\underset{\sim}{T} = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_s \end{bmatrix}$$ is as in Lemma 2.2.1. Then the joint pdf of $\underset{\sim}{u}' = (u_1, u_2, \ldots, u_s)$ is given by

$$f(u_1,u_2,\ldots,u_s) = [\Gamma(p/2)/\Gamma\{(p-s)/2\}]\ |C_2|^{-1/2}\pi^{-s/2}(1-u_2' C_2^{-1} u_2)^{(p-s-2)/2}$$

inside the region

$$u_2' C_2^{-1} u_2 \leq 1.$$

Proof : Let $S_2^2 = S_1^2 + \nu s_\nu^2$, where S_1^2 is as in Lemma 2.2.1. Since s_ν is an independent root mean square estimator of σ, hence T_2 and s_ν are independent. Consequently, by Lemma 2.2.1 T_2 and S_2^2 are independent, with

$$T_2 \overset{d}{=} N(0, C_2 \sigma^2), \text{ and}$$
$$S_2^2 \overset{d}{=} \sigma^2 \chi^2_{(p-s)}.$$

Without loss of generality we take $\sigma = 1$. Further

$$\begin{aligned} S_p^2 &= S^2 + \nu s_\nu^2 \\ &= S_1^2 + T_2' C_2^{-1} T_2 + \nu s_\nu^2 \\ &= S_2^2 + T_2' C_2^{-1} T_2. \end{aligned} \tag{2.2.6}$$

The joint pdf of $(u_1,u_2,\ldots,u_s)$ can now be obtained exactly as given by Ellenberg (1973), by considering suitable transformations. For this, we start with the joint density of T_2 and S_2^2 given by

$$\begin{aligned} f(t_1,t_2,\ldots,t_s, S_2^2) &= f_{T_2}(t_2)\ f_{S_2^2}(s_2^2) \\ &= \frac{|C_2|^{-1/2}}{(2\pi)^{s/2}} \exp\left[-(T_2' C_2^{-1} T_2)/2\right] \\ &\quad \cdot \frac{1}{2^{\nu_1}\,\Gamma(\nu_1)} (S_2^2)^{\nu_1 - 1} \exp(-S_2^2/2), \end{aligned}$$

where $\nu_1 = (p-s)/2$. Make a transformation

$$u_i = t_i/S_p, \quad i = 1,2,\ldots,s$$

and $$S_p = (S_2^2 + T_2' C_2^{-1} T_2)^{1/2}, \quad 0 < S_p < \infty.$$

This implies that $t_i = S_p u_i$ and

$$S_2^2 = S_p^2 - T_2' C_2^{-1} T_2 = S_p^2 - S_p^2 u_2' C_2^{-1} u_2 = S_p^2 (1-u_2' C_2^{-1} u_2).$$

The jacobian of the transformation is given by

$$|J_2| = \left|\frac{\partial(t_1,t_2,\ldots,t_s,S_2^2)}{\partial(u_1,u_2,\ldots,u_s,S_p)}\right| = \begin{vmatrix} S_p I_s & u_2 \\ -2S_p^2 u_2' C_2^{-1} & 2S_p(1-u_2' C_2^{-1} u_2) \end{vmatrix},$$

Now partitioning J_2 as

$$J_2 = \begin{bmatrix} J_{211} & J_{212} \\ J_{221} & J_{222} \end{bmatrix},$$

where $\underset{sxs}{J_{211}} = \underset{1x1}{S_p}\ \underset{sxs}{I_s}$, $\underset{sx1}{J_{212}} = \underset{sx1}{u_2}$,

$$\underset{1xs}{J_{221}} = -2\underset{1x1}{S_p^2}\ \underset{1xs}{u_2'}\ \underset{sxs}{C_2^{-1}} \quad \text{and} \quad \underset{1x1}{J_{222}} = 2\underset{1x1}{S_p}\ (1-\underset{1xs}{u_2'}\ \underset{sxs}{C_2^{-1}}\ \underset{sx1}{u_2}),$$

we get the determinant of J_2 as

$$\begin{aligned} |J_2| &= |J_{211}|\ |J_{222} - J_{221} J_{211}^{-1} J_{212}| \\ &= |S_p I_s|\ |2S_p(1-u_2' C_2^{-1} u_2) + 2 S_p^2 u_2' C_2^{-1} (1/S_p) u_2| \\ &= 2 S_p^{s+1}. \end{aligned}$$

Therefore,

$$f(u_1,u_2,\ldots,u_s/S_p) = \frac{|C_2|^{-1/2}\ 2^{-(\nu_1+s/2)}}{\pi^{s/2}\ G(\nu_1)}\ [s_p^2(1-u_2'\ C_2^{-1}\ u_2)]^{\nu_1-1}$$

$$\cdot\ e^{-s_p^2/2}\ \cdot\ (2s_p^{s+1})\cdot$$

Integrating out S_p from 0 to ∞, we get the joint pdf of $u_1,u_2,\ldots,u_s$ as

$$f(u_1,u_2,\ldots,u_s)$$

$$= \frac{|C_2|^{-1/2}\ 2^{-(\nu_1+s/2-1)}}{\pi^{s/2}\ G(\nu_1)}\ (1-u_2'\ C_2^{-1}\ u_2)^{\nu_1-1} \int_0^\infty S_p^{2\nu_1+s-1}\ e^{-s_p^2/2}\ d\,S_p$$

$$= \frac{G(\nu_1+s/2)\ |C_2|^{-1/2}}{\pi^{s/2}\ G(\nu_1)}\ (1-u_2'\ C_2^{-1}\ u_2)^{\nu_1-1}\ .$$

Substituting $\nu_1 = (p-s)/2$, we have

$$(2.2.7)\quad f(u_2) = [G(p/2)/G\{(p-s)/2\}]\ |C_2|^{-1/2}\ \pi^{-s/2}(1-u_2'C_2^{-1}u_2)^{(p-s-2)/2}.$$

This completes the proof of the theorem.

<u>Corollary 2.2.1</u>. For $s = 1$, the distribution of

$$u_1 = \sum_{i=1}^{n} m_{1i}\ w_i, \text{ is given by}$$

$$f(u_1) = \frac{1}{C^{1/2}\ B[1/2,(p-1)/2]}\ (1-u_1^2/C)^{(p-3)/2},\ u_1^2 \le C,$$

where $C = M_2\ R_2\ M_2'$, $M_2 = (m_{11},m_{12},\ldots,m_{1n})$ and $R_2 = ((\rho_{ij}))$.

<u>Proof</u> : For $s = 1$, we have from Lemma 2.2.1,

$$T_2 = t_1 = M_2\ Z_2 = \sum_{i=1}^{n} m_{1i}\ z_i = \sum_{i=1}^{n} m_{1i}\ e_i/\lambda_{ii}^{1/2}\ .$$

Hence $u_1 = t_1/S_p = \sum_{i=1}^{n} m_{1i}\, e_i/(S_p \lambda_{ii}^{1/2}) = \sum_{i=1}^{n} m_{1i} w_i$, from (1.2.5).

On applying Theorem 2.2.1, the pdf of u_1 is

$$f(u_1) = \frac{G(p/2)}{C^{1/2}\, \pi^{1/2}\, G\{(p-1)/2\}} (1-u^2/C)^{(p-3)/2}$$

$$= \frac{1}{C^{1/2}\, B[1/2,(p-1)/2]} (1-u^2/C)^{(p-3)/2}, \quad u_1^2 \leq C.$$

Hence the corollary is proved.

For the joint pdf of u_{ij} and $u_{i_1 j_1}$, we use equation (2.2.7) with $s = 2$ and take

$$\underset{2}{M} = \begin{bmatrix} 0 & 0 & \ldots\ldots\ldots\ldots\ldots & 1/[2(1+\rho_{ij})]^{1/2} & 0 & \ldots\ldots \\ 0 & 0 \ldots & 1/[2(1+\rho_{i_1 j_1})]^{1/2} & 0 & 0 & \ldots\ldots \end{bmatrix}$$

$$\begin{matrix} \ldots\ldots & 0 & 1/[2(1+\rho_{ij})]^{1/2} & 0 \\ \ldots\ldots & 1/[2(1+\rho_{i_1 j_1})]^{1/2} & 0 & \ldots\ldots\ 0 \end{matrix} \Bigg],$$

where the non-zero elements occur at ith and jth positions in the first row and i_1th and j_1th positions in the second row.

Now

$$\underset{2}{C} = \underset{2}{M}\, \underset{2}{R}\, \underset{2}{M}' = \begin{bmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{bmatrix},$$

where

$$(2.2.8) \quad \rho_1 = \rho_1 (u_{ij}, u_{i_1 j_1}) = \frac{\rho_{ii_1} + \rho_{ij_1} + \rho_{i_1 j} + \rho_{jj_1}}{2\,[(1+\rho_{ij})\,(1+\rho_{i_1 j_1})]^{1/2}}.$$

Consequently $|\underset{2}{C}| = (1 - \rho_1^2)$ and the joint distribution of u_{ij} and $u_{i_1 j_1}$ is given by

$$(2.2.9)\quad f(u_{ij},u_{i_1j_1}) = \frac{p-2}{2\pi(1-\rho_1^2)^{1/2}}\left[1-\frac{1}{1-\rho_1^2}\{u_{ij}^2+u_{i_1j_1}^2-2\rho_1 u_{ij}u_{i_1j_1}\}\right]^{(p-4)/2},$$

which is defined inside an ellipse $u_{ij}^2 + u_{i_1j_1}^2 - 2\rho_1 u_{ij} u_{i_1j_1} = 1 - \rho_1^2$. The parameter ρ_1 will be referred as the "shape" parameter of the joint pdf of u_{ij} and $u_{i_1j_1}$, because it determines the shape and orientation of this ellipse.

It is useful to note that $\rho_1(u_{ij},u_{i_1j_1})$ is identically equal to the product moment correlation coefficient between

$$z_{ij} = e_i/\lambda_{ii}^{1/2} + e_j/\lambda_{jj}^{1/2}$$

and $$z_{i_1j_1} = e_{i_1}/\lambda_{i_1i_1}^{1/2} + e_{j_1}/\lambda_{j_1j_1}^{1/2}.$$

To this end, note that with $\sigma = 1$

$$\text{Var}(z_{ij}) = 2(1+\rho_{ij}),$$

$$\text{Var}(z_{i_1j_1}) = 2(1+\rho_{i_1j_1})$$

and $$\text{Cov}(z_{ij},z_{i_1j_1}) = \rho_{ii_1} + \rho_{i_1j} + \rho_{ij_1} + \rho_{jj_1}$$

and the result follows immediately. This is useful in evaluating the shape parameter in special cases. It may also be noted that the joint pdf of u_{ij} and $u_{i_1j_1}$ is of the same form as that of w_1 and w_2 given at equation (2.2.1).

The bivariate distribution of two v_{ij}'s has exactly the

same form as (2.2.9), but in this case the $\underset{\sim}{M}_2$ matrix is taken as

$$\underset{\sim}{M}_2 = \begin{bmatrix} 0 & 0 & \cdots & 1/[2(1-\rho_{ij})]^{1/2} & 0 & \cdots \\ 0 & 0 & \cdots & 1/[2(1-\rho_{i_1 j_1})]^{1/2} & 0 & \cdots \\ \cdots & 0 & & 1/[2(1-\rho_{ij})]^{1/2} & 0 & \\ \cdots & 1/[2(1-\rho_{i_1 j_1})]^{1/2} & & 0 & \cdots & 0 \end{bmatrix},$$

the non-zero elements occur at the $\underline{i}$th and $\underline{j}$th positions in the first row and $\underline{i_1}$th and $\underline{j_1}$th positions in the second row. Further the shape parameter ρ_1' is given by

$$(2.2.10) \quad \rho_1' = \rho_1(v_{ij}, v_{i_1 j_1}) = \frac{\rho_{ii_1} - \rho_{ij_1} - \rho_{i_1 j} + \rho_{jj_1}}{2[(1-\rho_{ij})(1-\rho_{i_1 j_1})]^{1/2}} .$$

Again the shape parameter ρ_1 is equal to the product moment correlation between

$$(e_i/\lambda_{ii}^{1/2} - e_j/\lambda_{jj}^{1/2}) \text{ and } (e_{i_1}/\lambda_{i_1 i_1}^{1/2} - e_{j_1}/\lambda_{j_1 j_1}^{1/2}).$$

For s = 1, Corollary 2.2.1 immediately gives the univariate density functions of u_{ij} and v_{ij}, respectively, as

$$(2.2.11) \quad \begin{cases} f(u_{ij}) = (1-u_{ij}^2)^{(p-3)/2}/B[1/2,(p-1)/2] \ , -1 \le u_{ij} \le 1, \\ f(v_{ij}) = (1-v_{ij}^2)^{(p-3)/2}/B[1/2,(p-1)/2], \ -1 \le v_{ij} \le 1. \end{cases}$$

Remark :

Margolin (1977) has obtained the Ellenberg's result for

the joint density of $(w_1, w_2, \ldots, w_s)$. He has used the following theorem with its corollary for establishing this result.

Theorem. Consider a set of n random variables $\underset{\sim}{Z}^* = (Z_1^*, Z_2^*, \ldots, Z_n^*)$ whose distribution function has only one parameter $\theta > 0$. Assume further that

(a) $T^* = u(\underset{\sim}{Z}^*)$ is sufficient for θ, and T^* has a gamma distribution $G(d,\theta)$ for $d > 0$, that is

$$f_{T^*}(t^*) = \theta^d \, t^{*d-1} \, e^{-\theta t^*}/G(d), \quad t^* \geq 0.$$

(b) H is a function from R^n to R such that $E_{\underset{\sim}{Z}^*}\{|H(\underset{\sim}{Z}^*)|\}$ exists and is finite for all $\theta > 0$; and

(c) $\underset{\sim}{S}^* = (S_1^*, S_2^*, \ldots, S_n^*)$ is a vector of studentized analogues of $\underset{\sim}{Z}^*$, namely for certain strictly positive functions $h_i(T^*) = 1$, assume that $S_i^* = Z_i^*/h_i(T^*)$ $(i = 1, \ldots, n)$. Then it follows that

$$E_{\underset{\sim}{S}^*}\{H(\underset{\sim}{S}^*)\} = G(d) \; \mathcal{L}^{-1}\left[\theta^{-d} \, E_{\underset{\sim}{Z}^*}\{H(\underset{\sim}{Z}^*)\}; \, 1\right],$$

where $\mathcal{L}^{-1}$ is the inverse of Laplace transform, that is, if $\mathcal{L}\{f(t^*);\theta\} = g(\theta) = \int_0^\infty e^{-\theta} \, f(t^*)\,dt^*$, then the inverse Laplace transform of $g(\theta)$ is $\mathcal{L}^{-1}\{g(\theta); t^*\}$.

Corollary. Assume the conditions of the theorem. In addition to that, if the probability density functions $f_{\underset{\sim}{S}^*}$ and $f_{\underset{\sim}{Z}^*}$ exists, then

$$(2.2.12) \qquad f_{\underset{\sim}{S}^*}(\underset{\sim}{s}^*) = G(d) \; \mathcal{L}^{-1}\{\theta^{-d} \, f_{\underset{\sim}{Z}^*}(\underset{\sim}{s}^*, \theta); 1\}.$$

Using these, Margolin has obtained the marginal pdf of $\underset{\sim}{w} = (w_1, w_2, \ldots, w_s)$ as

$$(2.2.13)\quad f_{\underset{\sim}{W}}(\underset{\sim}{w}) = [\,G\{(n-k+\nu)/2\}/G\{(n-k+\nu-s)/2\}]\ |\underset{\sim}{R}|^{-1/2}\ \pi^{-s/2}\ (1-\underset{\sim}{w}'\ \underset{\sim}{R}^{-1}\ \underset{\sim}{w})^{(n-k+\nu-s-2)/2},$$

which is same as derived by Ellenberg (1973).

Possibly, Margolin's results are valid under weaker conditions, since he implicitly assumes that the residual sum of squares S^2 is sufficient for σ^2, which is not true. Using the same argument and applying the results in our case, we get the joint pdf of $\underset{\sim}{u}$ exactly same as derived in equation (2.2.7).

2.3. Nominal percentile points

The result used in the following lemma is used at several places in this thesis.

Lemma 2.3.1.

Let $z_1, z_2, \ldots, z_N$ be N identically distributed random variables with common pdf

$$p(z_1) = (1 - z_1^2)^{(p-3)/2}/B(\tfrac{1}{2}, \tfrac{p-1}{2}), \quad -1 \le z_1 \le 1,$$

and $Z = \text{Max}\{z_i : i = 1,2,\ldots,N\}$. Then an upper limit z_α for the exact percentile point $Z_{\alpha(e)}$ of Z at α level of significance for $(\alpha/N) \le 0.5$ is given by the equation

$$I_{1-z_\alpha^2}[(p-1)/2,\ 1/2] = 2\alpha/N,$$

where $I_x(a,b) = \frac{1}{B(a,b)} \int_0^x t^{a-1}(1-t)^{b-1}\, dt$ is the incomplete beta function. The upper limit z_α will be called a nominal upper 100α percent critical point of Z.

Proof: Clearly

$$Pr(Z > z) = Pr(\underset{i}{\text{Max}}\ z_i > z) = Pr\left(\bigcup_{i=1}^{N} (z_i > z)\right).$$

Applying the first Bonferroni inequality, we get

$$Pr(Z > z) \leq N\ Pr(z_1 > z).$$

Let z_α be the solution of

$$N\ Pr(z_1 > z) = \alpha\ .$$

Then we immediately get

$$Pr(Z > z_\alpha) \leq \alpha\ ,$$

and z_α as an upper limit for $Z_{\alpha(e)}$. Thus, for getting z_α, we solve

$Pr(z_1 > z_\alpha) = \frac{\alpha}{N}$. But for $\frac{\alpha}{N} \leq 0.5$ we have $z_\alpha \geq 0$ and

$$Pr(z_1 > z_\alpha) = \int_{z_\alpha}^{1} (1-z_1^2)^{(p-3)/2}\, dz_1 / B\left[\frac{1}{2}\ ,\ \frac{(p-1)}{2}\right]$$

$$= \int_{z_\alpha^2}^{1} t^{-\frac{1}{2}} (1-t)^{(p-3)/2}\, dt / \{2B(\tfrac{1}{2},\ \tfrac{p-1}{2})\}$$

$$= \frac{1}{2} I_{1-z_\alpha^2}(\tfrac{p-1}{2}\ ,\ \tfrac{1}{2}).$$

Consequently, $\sqrt{}$ solution of

$$I_{1-z_\alpha^2}[(p-1)/2, 1/2] = 2\alpha/N$$

gives the desired upper limit for $Z_{\alpha(e)}$.

Corollary 2.3.1. If $Z_1 = \underset{1\leq i\leq n}{\text{Max}} \{|z_i|\}$, then an upper limit $z_{1\alpha}$ for the true 100α percent point of Z_1 is given by

$$I_{1-z_{1\alpha}^2}[(p-1)/2,\ 1/2] = \alpha/N .$$

Proof : As before, we have

$$Pr(Z_1 > z_{1\alpha}) \leq N\ Pr(|z_1| > z_{1\alpha}) = 2N\ Pr(z_1 > z_{1\alpha}),$$

and the result follows. Note that $z_{1\alpha} \geq 0$ since $N \geq 1$, and we do not require that $z_{1\alpha} \geq 0$.

Upper and lower limits for the true percentage points can be obtained by using the Bonferroni (David, 1956) and other inequalities.

Since $U = \underset{1\leq i<j\leq n}{\text{Max}} u_{ij}$, an upper bound for true percentile point at α level of significance can be obtained by using Lemma 2.3.1. With $N = \binom{n}{2}$, we immediately get

$$(2.3.1) \qquad I_{1-u_\alpha^2}[(p-1)/2, 1/2] = 4\alpha/[n(n-1)] .$$

Solution of equation (2.3.1) gives nominal upper percentage point u_α. This can be obtained either from the tables of incomplete beta function prepared by Pearson (1968) or by the method described in Appendix I.

Similarly for v_α the equation to be solved is

$$(2.3.2) \qquad I_{1-v_\alpha^2}[(p-1)/2,\ 1/2] = 2\alpha/[n(n-1)] .$$

Table 2.3.1 gives nominal upper critical values u_α for α = 0.005, 0.01, 0.025, 0.05 and 0.10; n = 5(1)12, 14(1)16(2) 20,21,24,25,27,28(2) 32,33,35,36,40,42,45,48(1)50(5) 60(10)100 and for k = 1(1) min (n-2,15). The quantity on the R.H.S. of (2.3.1), that is $4\alpha/[n(n-1)]$ is calculated first. Then using the inverse of incomplete beta function, which is given in the procedure described in Appendix I, the values of $1-u_\alpha^2$ and of u_α are determined. The values of n for which these calculations are done are chosen such that this table would give critical values for two-way tables with r rows and c columns, for $r+c \leq 14$. Additional values of n are also included. The same table can be used for v_α also for α = 0.01, 0.02, 0.05, 0.10 and 0.20.

For comparing the performance of tests based on U and V with sequential test, a brief table of nominal upper critical values u_α for (n,k) = (10,1), (11,1), (20,8), (21,1) and (48,13); and α = 0.02625 is given in Table 2.3.2. For same α, and (n,k) = (20,1), (20,8), (48,13); the v_α values are given in Table 2.3.3.

As an example for the use of these critical values, we again consider Example 2.1.1. There we calculated the value of U, V and U_1 as 0.6788, 0.8617 and 0.8022 respectively. From Table 2.3.1 we find the u_α value for α = 0.05; n = 20; k = 8 is 0.8244 and the v_α value for the same n,k and α is 0.8465. Thus only V exceeds the tabulated value at 5 percent level of

significance and hence we decide that there are two outliers on either side.

In general, it is extremely difficult to formulate a model that there are two outliers on right, or on left, or one on each side. Consequently, we recommend that all these three statistics U, U_1 and V should be examined for possible outliers. A decision to declare outliers can then be based on observed significance probability (P-value). For the present example, the value of U is considerably small, and hence we calculate P-values for U_1 and V only, which are equal to 0.092 and 0.029 respectively. This shows that there are two outliers, one on each side and we accordingly declare y_{43} and y_{44} as the two outliers. It may be noted that using a two-sided statistic for the detection of single outlier at a significance level $\alpha = 0.05$, Doornbos (1981) came to the conclusion that y_{43} is an outlying observation.

A lower limit $u_{*\alpha}$ can be obtained by considering the second Bonferroni inequality and solving for $u_{*\alpha}$ in

$$(2.3.3) \quad \binom{n}{2} \Pr(u_{ij} > u_{*\alpha}) - \Sigma\Sigma \Pr(u_{ij} > u_{*\alpha},\ u_{i_1 j_1} > u_{*\alpha}) = \alpha,$$

where the double sum is over all distinct terms, which appear in the second term of Bonferroni inequality. Note that equation (2.3.3) may not have a solution for all values of n and α, especially for large values of these quantities. Solution of (2.3.3) involves the calculation, of bivariate probabilities,

which among other things depend on the shape parameter $\rho = \rho_1(u_{ij}, u_{i_1 j_1})$. We discuss this evaluation in next section.

2.4. Evaluation of bivariate probabilities

An expression for the bivariate probability $Pr(w_1 > h, w_2 > k)$ has been obtained by Joshi (1972,1975) where (w_1, w_2) follow the joint distribution given at equation (2.2.1). Here we intend to find a recurrence relation involving p for this function. Since the joint pdf of $(u_{ij}, u_{i_1 j_1})$ given at equation (2.2.9) is analogous to that of (w_1, w_2), hence we continue to use the simpler notation of w_1, w_2 etc. involving only one suffix. Let

$$M(h,k,\rho,p) = Pr(w_1 > h,\ w_2 > k)$$

$$(2.4.1) \qquad = \frac{(p-2)}{2\pi(1-\rho^2)^{1/2}} \iint \left(1 - \frac{w_1^2 - 2\rho\, w_1 w_2 + w_2^2}{1-\rho^2}\right)^{(p-4)/2} dw_1\, dw_2$$

where the region of integration is given by

$$(2.4.2) \qquad w_1 > h,\ w_2 > k,\ w_1^2 - 2\rho\, w_1 w_2 + w_2^2 \leq 1 - \rho^2.$$

Define

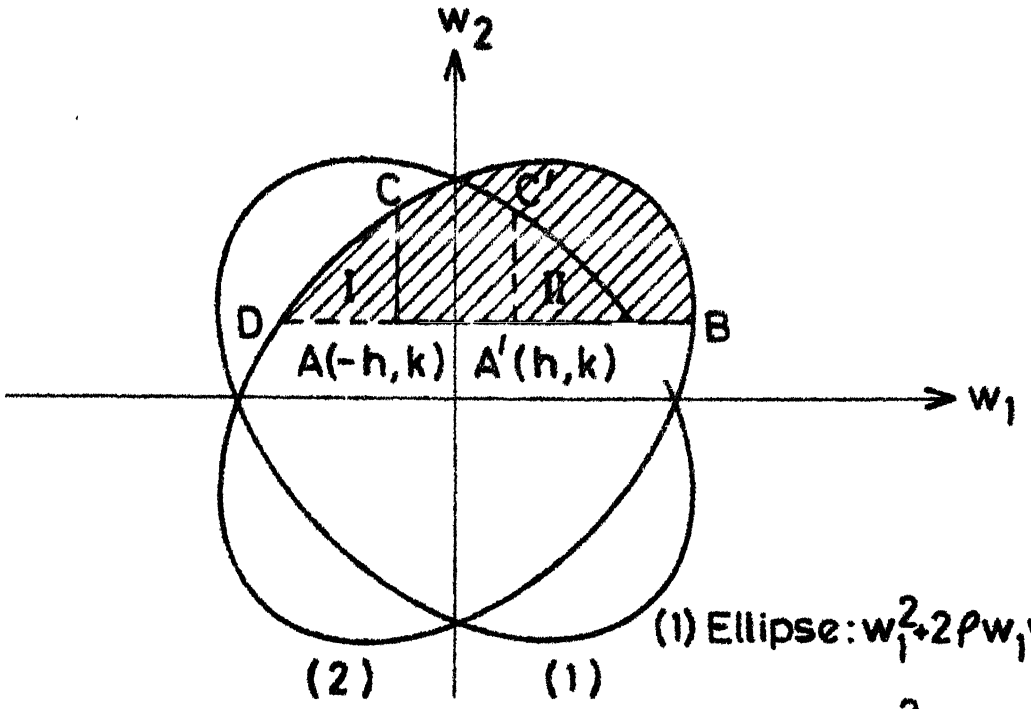
$$Q(a) = Pr(w_1 > a) = \int_a^1 (1-w_1^2)^{(p-3)/2} dw_1 / B\left[1/2, (p-1)/2\right].$$

Then the M function has the following properties :

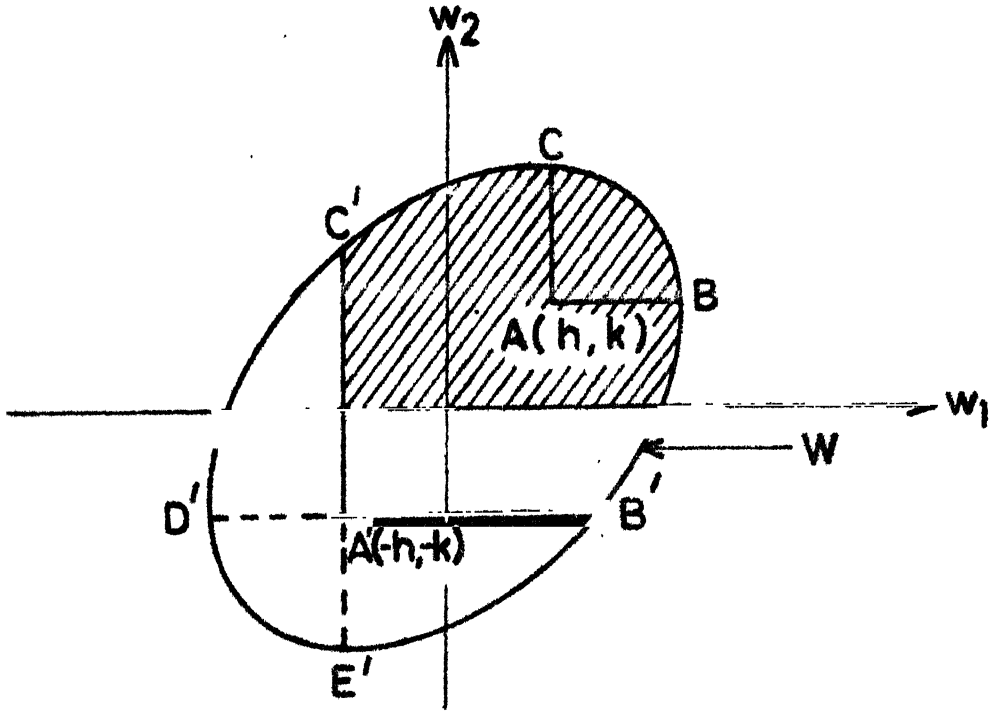
(I) $M(h,k,\rho,p) = M(k,h,\rho,p)$.

This follows, because of the symmetry of w_1 and w_2 in equation (2.4.1).

(II) $M(-h,k,\rho,p) + M(h,k,-\rho,p) = Q(k)$.



.2.4.1. Showing the regions for M(h,k, ρ,p) and M(h,k,- ρ,p).



lg. 2.4.2. Showing the regions for M(h,k, ρ,p) and M(-h,-k, ρ,p).

This follows from the figure 2.4.1. $Q(k)$ represents the probability of a point in the region covered by the segment DB and the part of the ellipse DCB. $M(-h,k,\rho,p)$ is the probability of a point lying in the region CABC. Also by symmetry we have, region DACD = region $A'B'C'A'$. Hence the result follows.

(III) $M(-h,-k,\rho,p) = 1-Q(h) - Q(k) + M(h,k,\rho,p)$.

This follows from figure 2.4.2. We see that $M(-h,-k,\rho,p)$ is the probability of the region covered by $C'A'B'BCC'$. Suppose W is the whole region covered by the ellipse, then

$$C'A'B'BCC' = W-E'D'C'E' - D'B'E'D' + E'D'A'E'$$

$$= W-E'D'C'E' - D'B'E'D' + BACB,$$

since $E'D'A'E' = BACB$, by symmetry. Hence, on integrating over these regions, we get the required result. Mathematical proofs of these results can also be derived easily.

From these relations it is obvious that we have to consider only $h,k \geq 0$. Let A be the point (h,k) in the (w_1,w_2)-plane. Without loss of generality we take A to be inside the ellipse

$$(2.4.3) \qquad w_1^2 - 2\rho\, w_1 w_2 + w_2^2 = 1 - \rho^2,$$

otherwise the required probability is zero. The region of integration is then the shaded area ABCA (see figure 2.4.3).

Theorem 2.4.1. Consider the figure 2.4.3.

Let

$$M(h,k,\rho,p) = Pr\,(w_1 > h,\ w_2 > k) = Pr\,[(w_1,w_2) \in ABCA],$$

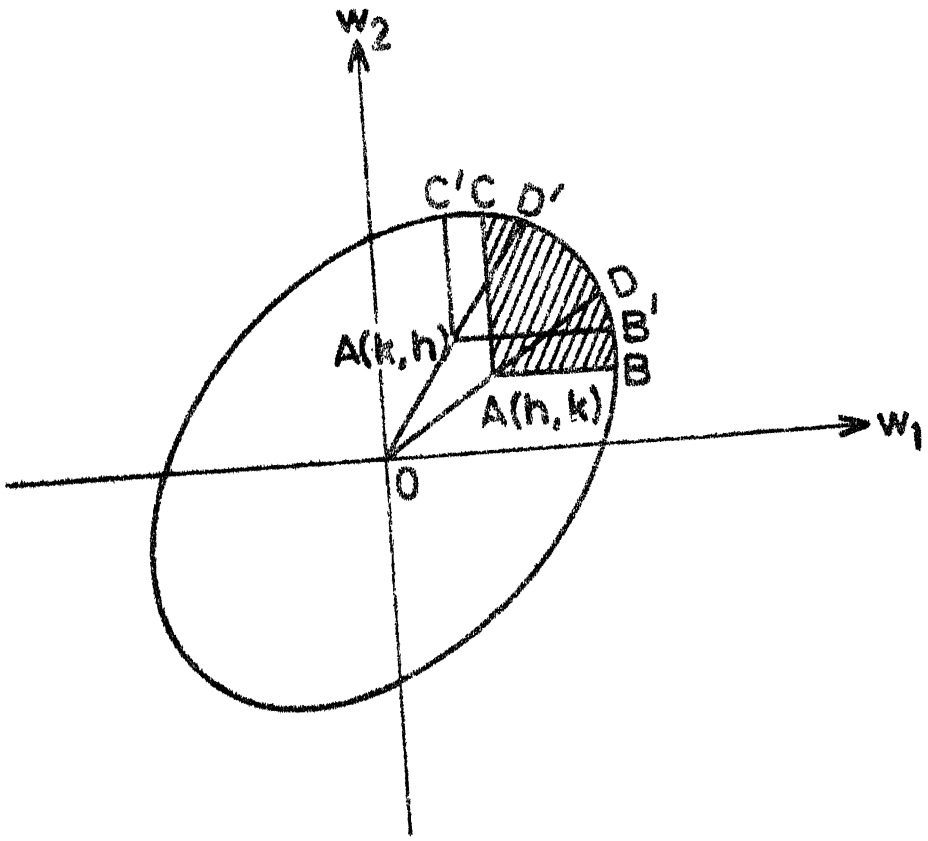


Fig. 2.4.3. Showing the regions for $M_1(h,k,P,p)$ and $M(h,k,P,p)$

$M_1(h,k,\rho,p) = \Pr[(w_1,w_2) \in ABDA]$,

where AD is the line segment passing through the origin and the point (h,k) and intersecting the ellipse at the point D.

Then for all $h,k \geq 0$,

(i) $M(h,k,\rho,p) = M_1(h,k,\rho,p) + M_1(k,h,\rho,p)$,

(ii) $M_1(\rho k, k,\rho,p) = \frac{1}{4} I_{1-k^2}[(p-1)/2, 1/2]$,

(iii) $M_1(h,k,\rho,p) = M_1(\rho k,k,\rho,p) - \text{sign}(h-\rho k)L(k,c,p)$,

where sign (u) = +1 or -1 according as $u > 0$ or $u < 0$ respectively,

(2.4.4) $c = |h-\rho k|/(h^2+k^2-2\rho hk)^{1/2}$

and

(2.4.5) $L(k,c,p) = \frac{1}{2\pi}\int_0^c \frac{1}{(1-z^2)^{1/2}}\left(1 - \frac{k^2}{1-z^2}\right)^{(p-2)/2} dz$.

Proof : (i) This is an obvious result and is proved in Joshi (1975).

(ii) Joshi (1975) has also obtained expressions for $M(h,k,\rho,p)$ and $M_1(h,k,\rho,p)$ as

(2.4.6) $M(h,k,\rho,p)$

$$= \frac{1}{2\pi}\int_{hk-(1-h^2)^{1/2}(1-k^2)^{1/2}}^{\rho} \frac{1}{(1-z^2)^{1/2}}\left(1 - \frac{h^2+k^2-2zhk}{1-z^2}\right)^{(p-2)/2} dz$$

and

(2.4.7) $M_1(h,k,\rho,p) =$

$$= \frac{1}{2\pi}\int_{\arctan[k/(1-k^2)^{1/2}]}^{\arctan[k(1-\rho^2)^{1/2}/(h-\rho k)]} (1-k^2\,\text{cosec}^2\theta)^{(p-2)/2}\, d\theta.$$

Substituting $z = -\cos\theta$ in (2.4.7), we get

$$(2.4.8)\quad M_1(h,k,\rho,p) = \int_{-(1-k^2)^{1/2}}^{-(h-\rho k)/(h^2+k^2-2\rho hk)^{1/2}} \frac{1}{2\pi(1-z^2)^{1/2}} \left(1 - \frac{k^2}{1-z^2}\right)^{(p-2)/2} dz .$$

Similarly, on interchanging the roles of h and k, we obtain $M_1(k,h,\rho,p)$.

When $h-\rho k = 0$, that is, $h = \rho k$, equation (2.4.8) gives

$$M_1(h,k,\rho,p) = M_1(\rho k,k,\rho,p)$$

$$= \frac{1}{2\pi} \int_{-(1-k^2)^{1/2}}^{0} \frac{1}{(1-z^2)^{1/2}} \left(1 - \frac{k^2}{1-z^2}\right)^{(p-2)/2} dz$$

$$= M(0,k,0,p), \text{ from equation (2.4.6)}$$

$$= Pr(w_1 > 0,\ w_2 > k \mid \rho = 0, p)$$

$$= \frac{p-2}{2\pi} \int_k^1 \left[\int_0^{(1-w_2^2)^{1/2}} (1-w_1^2-w_2^2)^{(p-4)/2}\, dw_1 \right] dw_2 ,$$

where the last equality follows by considering the joint pdf of w_1 and w_2 for $\rho = 0$ and integrating over the region $\{w_1 > 0, w_2 > k\}$.

Letting $t = w_1/(1-w_2^2)^{1/2}$,

$$M_1(\rho k,k,\rho,p) = \frac{p-2}{2\pi} \int_k^1 \left[\int_0^1 (1-w_2^2)^{(p-3)/2} (1-t^2)^{(p-4)/2}\, dt \right] dw_2$$

$$= \frac{1}{2} \int_k^1 (1-w_2^2)^{(p-3)/2}\, dw_2 / B\left[1/2, (p-1)/2\right]$$

$$= \frac{1}{4} I_{1-k^2}\left[(p-1)/2,\ 1/2\right] .$$

(iii) When $h-\rho k < 0$, then the upper limit of integral appearing on the R.H.S. of (2.4.8) is c, with c given at equation (2.1.4). Hence $c > 0$ and

$$M_1(h,k,\rho,p) = \frac{1}{2\pi} \int_{-(1-k^2)^{1/2}}^{c} \frac{1}{(1-z^2)^{1/2}} \left(1 - \frac{k^2}{1-z^2}\right)^{(p-2)/2} dz$$

$$= \frac{1}{2\pi} \int_{-(1-k^2)^{1/2}}^{0} \frac{1}{(1-z^2)^{1/2}} \left(1 - \frac{k^2}{1-z^2}\right)^{(p-2)/2} dz$$

$$+ \frac{1}{2\pi} \int_{0}^{c} \frac{1}{(1-z^2)^{1/2}} \left(1 - \frac{k^2}{1-z^2}\right)^{(p-2)/2} dz.$$

From the result proved at (ii) above the first term on the R.H.S. is equal to $M_1(\rho k,k,\rho,p)$. Consequently,

$$M_1(h,k,\rho,p) = M_1(\rho k,k,\rho,p) + L(k,c,p),$$

where L(k,c,p) is given at (2.4.5).

Similarly, when $h-\rho k > 0$, we get

$$M_1(h,k,\rho,p) = M_1(\rho k,k,\rho,p) - L(k,c,p).$$

This completes the proof of the theorem.

This theorem shows that in order to evaluate $M_1(h,k,\rho,p)$ and hence $M(h,k,\rho,p)$, we need to evaluate L(k,c,p), since the incomplete beta functions appearing in these expressions are extensively tabulated or can be obtained by the method described in Appendix I. For the evaluation of L(k,c,p), we derive a recursive relation in the Theorem 2.4.2.

Theorem 2.4.2. Let k and c be positive and

$$L(k,c,p) = \frac{1}{2\pi} \int_0^c \frac{1}{(1-z^2)^{1/2}} \left(1 - \frac{k^2}{1-z^2}\right)^{(p-2)/2} dz,$$

then

$$L(k,c,2) = \frac{1}{2\pi} \operatorname{Sin}^{-1}(c)$$

$$L(k,c,3) = \frac{1}{4\pi} \left[\operatorname{Sin}^{-1}\left(\frac{k^2+2c^2-1}{1-k^2}\right) - k \operatorname{Sin}^{-1}\left\{\frac{2k^2-(1+k^2)(1-c^2)}{(1-k^2)(1-c^2)}\right\}\right] + \frac{1-k}{8}$$

and for $p > 3$,

$$L(k,c,p) = L(k,c,p-2) - \frac{k(1-k^2)^{(p-3)/2}}{4\pi} B\left(\frac{p-2}{2}, \frac{1}{2}\right) \cdot I_{c^2k^2/[(1-c^2)(1-k^2)]}\left[1/2,(p-2)/2\right].$$

Proof : For $p = 2$, we have

$$L(k,c,2) = \frac{1}{2\pi} \int_0^c \frac{1}{(1-z^2)^{1/2}} dz = \frac{1}{2\pi} \operatorname{Sin}^{-1}(c).$$

Next consider the case $p = 3$, that is

$$L(k,c,3) = \frac{1}{2\pi} \int_0^c \frac{1}{(1-z^2)^{1/2}} \left(1 - \frac{k^2}{1-z^2}\right)^{1/2} dz.$$

Putting $z_1 = k^2/(1-z^2)$, we get $z = [(z_1-k^2)/z_1]^{1/2}$

and $dz = k^2\, dz_1/[2z_1^{3/2}(z_1-k^2)^{1/2}]$. Consequently

$$L(k,c,3) = \frac{k}{4\pi} \int_{k^2}^{k^2/(1-c^2)} \frac{(1-z_1)\, dz_1}{z_1(1-z_1)^{1/2}\,(z_1-k^2)^{1/2}}$$

$$= \frac{k}{4\pi} \int_{k^2}^{k^2/(1-c^2)} \frac{(1-z_1)\, dz_1}{z_1(-z_1^2+(1+k^2)\, z_1-k^2)^{1/2}}$$

$$= \frac{k}{4\pi} \int_{k^2}^{k^2/(1-c^2)} \frac{dz_1}{z_1\, Z^{1/2}} - \frac{k}{4\pi} \int_{k^2}^{k^2/(1-c^2)} \frac{dz_1}{Z^{1/2}},$$

where $Z = a+bz_1 + dz_1^2$ with $a = -k^2$, $b = 1+k^2$ and $d = -1$.

Using the tables of standard integrals (for example, see Selby and Girling (1965), p. 307), we get

$$L(k,c,3) = \frac{k}{4\pi} \frac{1}{(-a)^{1/2}} \operatorname{Sin}^{-1} \left[\frac{bz_1+2a}{z_1(b^2-4ad)^{1/2}}\right] \Bigg|_{k^2}^{k^2/(1-c^2)}$$

$$- \frac{k}{4\pi} \frac{1}{(-d)^{1/2}} \operatorname{Sin}^{-1} \left[\frac{-2dz_1-b}{(b^2-4ad)^{1/2}}\right] \Bigg|_{k^2}^{k^2/(1-c^2)}.$$

That is,

$$L(k,c,3) = \frac{1}{4\pi} \left[\operatorname{Sin}^{-1} \left(\frac{k^2+2c^2-1}{1-k^2}\right) - k \operatorname{Sin}^{-1} \left\{\frac{2k^2-(1+k^2)(1-c^2)}{(1-k^2)(1-c^2)}\right\}\right] + \frac{1-k}{8}.$$

Now for $p > 3$,

$$L(k,c,p) = \frac{1}{2\pi} \int_0^c \frac{1}{(1-z^2)^{1/2}} \left(1 - \frac{k^2}{1-z^2}\right)^{(p-2)/2} dz$$

$$= \frac{1}{2\pi} \int_0^c \frac{1}{(1-z^2)^{1/2}} \left(1 - \frac{k^2}{1-z^2}\right)^{(p-4)/2} dz$$

$$- \frac{k^2}{2\pi} \int_0^c \frac{1}{(1-z^2)^{3/2}} \left(1 - \frac{k^2}{1-z^2}\right)^{(p-4)/2} dz$$

$$= L(k,c,p-2) - I,$$

where the integral I is given by

$$(2.4.9) \qquad I = \frac{k^2}{2\pi} \int_0^c \frac{1}{(1-z^2)^{3/2}} \left(1 - \frac{k^2}{1-z^2}\right)^{(p-4)/2} dz.$$

Putting $z_1 = k^2/(1-z^2)$, we get

$$I = \frac{k^2}{2\pi} \int_{k^2}^{k^2/(1-c^2)} \frac{z_1^{3/2}}{k^3} (1-z_1)^{(p-4)/2} \frac{k^2\, dz_1}{2 z_1^{3/2} (z_1-k^2)^{1/2}}$$

or

$$I = \frac{k}{4\pi} \int_{k^2}^{k^2/(1-c^2)} (1-z_1)^{(p-4)/2} (z_1-k^2)^{-1/2} dz_1 .$$

Letting $z_1 - k^2 = z_2 (1-k^2)$, we have

$$dz_1 = (1-k^2)\, dz_2,\ 1-z_1 = (1-z_2)(1-k^2), \text{ and}$$

$$I = \frac{k(1-k^2)^{(p-3)/2}}{4\pi} \int_0^{c^2k^2/[(1-c^2)(1-k^2)]} (1-z_2)^{(p-4)/2} z_2^{-1/2} dz_2$$

$$= \frac{k(1-k^2)^{(p-3)/2}}{4\pi} B\left(\frac{p-2}{2}, \frac{1}{2}\right) I_{\frac{c^2k^2}{(1-c^2)(1-k^2)}}\left(\frac{1}{2}, \frac{p-2}{2}\right).$$

This completes the proof of the theorem.

Note that

$$I_{a^2}(1/2, 1/2) = \frac{2}{\pi} \operatorname{Sin}^{-1}(a).$$

Consequently,

$$\operatorname{Sin}^{-1}(a) = \begin{cases} \frac{\pi}{2} I_{a^2}(1/2, 1/2) & \text{if } a > 0 \\ -\frac{\pi}{2} I_{a^2}(1/2, 1/2) & \text{if } a < 0 \end{cases}$$

and it is possible to evaluate L(k,c,p) function of Theorem 2.4.2

in terms of incomplete beta function for all values of p.

The bivariate probability can also be evaluated by numerical integration from equation (2.4.6). But for smaller values of p this recursive method, which essentially involves the evaluation of incomplete beta integrals is more accurate.

Results for the case $h = k > 0$.

We next discuss the important case $h = k > 0$. This is useful in our application as we need bivariate probabilities like $\Pr(u_{ij} > h,\ u_{i_1 j_1} > h) \equiv M(h,h,\rho_1,p)$, where ρ_1 is the shape parameter given at equation (2.2.8). Similarly

$$(2.4.10)\quad \Pr(|v_{ij}| > h,\ |v_{i_1 j_1}| > h) \equiv 2[M(h,h,\rho_1',p)+M(h,h,-\rho_1',p)]$$

where ρ_1' is given at equation (2.2.10). This is required for evaluating lower bound. Since the joint pdf of $(u_{ij},u_{i_1 j_1})$ is identical as that of (w_1,w_2) given at equation (2.2.1), hence using results due to Srikantan (1961), Joshi (1972) etc. we have

$$(2.4.11)\quad \Pr(w_1 > h,\ w_2 > h \mid \rho,p) = M(h,h,\rho,p) = 0$$

whenever $h \geq \{(1+\rho)/2\}^{1/2}$, that is, whenever

$$(2.4.12)\qquad \rho \leq 2h^2 - 1.$$

Similarly,

$$(2.4.13)\quad \Pr(|w_1| > h,\ |w_2| > h) = 0 \text{ whenever } h \geq \{(1+|\rho|)/2\}^{1/2}.$$

Equation (2.4.11) shows that $M(h,h,\rho,p) = 0$ for all $h \geq \{(1+\rho)/2\}^{1/2}$. For other non-negative values of h, the following theorem gives a systematic evaluation of $M(h,h,\rho,p)$ in terms of incomplete beta functions.

Theorem 2.4.3. For $0 \leq h < \{(1+\rho)/2\}^{1/2}$, $M(h,h,\rho,p)$ is given by the following relations :

(i) $M(h,h,\rho,2) = [\mathrm{Sin}^{-1}\rho - \mathrm{Sin}^{-1}(2h^2-1)]/2\pi$

(ii) $M(h,h,\rho,3) = \frac{1-h}{4} - \frac{1}{2\pi}\left[\mathrm{Sin}^{-1}\frac{h^2-\rho}{1-h^2} - h\,\mathrm{Sin}^{-1}\frac{4h^2-(1+h^2)(1+\rho)}{(1+\rho)(1-h^2)}\right]$

(iii) For $p > 3$,

$$M(h,h,\rho,p) = M(h,h,\rho,p-2) - \frac{h(1-h^2)^{(p-3)/2}}{2\pi} B\left(\frac{p-2}{2}, \frac{1}{2}\right) I_{\frac{1+\rho-2h^2}{(1+\rho)(1-h^2)}}\left(\frac{p-2}{2}, \frac{1}{2}\right).$$

Proof : From equation (2.4.6), we have

$$(2.4.14)\quad M(h,h,\rho,p) = \frac{1}{2\pi}\int_{2h^2-1}^{\rho} \frac{1}{(1-z^2)^{1/2}}\left(1-\frac{2h^2}{1+z}\right)^{(p-2)/2} dz .$$

For $p = 2$, it immediately gives

$$M(h,h,\rho,2) = \frac{1}{2\pi}\int_{2h^2-1}^{\rho} \frac{1}{(1-z^2)^{1/2}}\,dz = \frac{1}{2\pi}\,\mathrm{Sin}^{-1}(z)\Big|_{2h^2-1}^{\rho}$$

$$= \frac{1}{2\pi}\left[\mathrm{Sin}^{-1}(\rho) - \mathrm{Sin}^{-1}(2h^2-1)\right].$$

For $p \geq 3$, substituting

$$z_1 = 2h^2/(1+z)$$

in equation (2.4.14) and simplifying, we have

$$(2.4.15)\quad M(h,h,\rho,p) = \frac{h}{2\pi}\int_{2h^2/(1+\rho)}^{1} \frac{1}{z_1(z_1-h^2)^{1/2}}(1-z_1)^{(p-2)/2}\,dz_1.$$

For $p = 3$, equation (2.4.11) gives

$$M(h,h,\rho,p) = \frac{h}{2\pi}\int_{2h^2/(1+\rho)}^{1} \frac{(1-z_1)^{1/2}}{z_1(z_1-h^2)^{1/2}}\,dz_1.$$

Multiplying the numerator and denominator of the integrand by $(1-z_1)^{1/2}$ and proceeding as for $L(k,c,3)$ in Theorem 2.4.2, the result for $p = 3$ follows. Next for $p > 3$, writing $(1-z_1)^{(p-2)/2}$ as $(1-z_1)(1-z_1)^{(p-4)/2}$ in equation (2.4.15), we get

$$(2.4.16)\quad M(h,h,\rho,p) = M(h,h,\rho,p-2) - I,$$

where

$$I = \frac{h}{2\pi}\int_{2h^2/(1+\rho)}^{1} \frac{1}{(z_1-h^2)^{1/2}}(1-z_1)^{(p-4)/2}\,dz_1.$$

Substituting

$z_1-h^2 = z_2(1-h^2)$, we get

$$I = \frac{h}{2\pi}(1-h^2)^{(p-3)/2}\int_{h^2(1-\rho)/[(1-h^2)(1+\rho)]}^{1} z_2^{1/2}(1-z_2)^{(p-4)/2}\,dz_2$$

$$= \frac{h}{2\pi}(1-h^2)^{(p-3)/2}\, I_{\frac{1+\rho-2h^2}{(1-h^2)(1+\rho)}}\left(\frac{p-2}{2},\frac{1}{2}\right) B\left(\frac{p-2}{2},\frac{1}{2}\right).$$

Substituting in equation (2.4.16) the result follows.

2.5. Bounds for bivariate probabilities

For $\rho \leq 0$, and c_1 and c_2 of the same sign, Joshi (1972) has shown that

$$(2.5.1) \qquad \Pr(w_1 \leq c_1,\ w_2 \leq c_2) \leq \prod_{i=1}^{2} \Pr(w_i \leq c_i).$$

Since the u_{ij}'s and v_{ij}'s also have exactly the same distribution as w_i's, hence (2.5.1) will hold for them.

When $c_1 = c_2 = c$ is positive (2.5.1) is equivalent to

$$\Pr(w_1 > c,\ w_2 > c) \leq \prod_{i=1}^{2} \Pr(w_i > c) = [\Pr(w_1 > c)]^2 .$$

Since

$$\Pr(w_i > c) = \frac{1}{2} I_{1-c^2}\left(\frac{p-1}{2}, \frac{1}{2}\right),$$

hence an upper bound of the bivariate probability when $\rho \leq 0$ is given by

$$(2.5.2) \qquad \Pr(w_1 > c,\ w_2 > c) \leq \left[\frac{1}{2} I_{1-c^2}\left(\frac{p-1}{2}, \frac{1}{2}\right)\right]^2 .$$

For all values of ρ, we can use the following inequality for obtaining a simple bound for the bivariate probability. The argument is essentially same as that of Cook and Prescott (1981). We have

$$\Pr(w_1 > c,\ w_2 > c) \leq \Pr(w_1 + w_2 \geq 2c)$$

$$= \Pr\left[\frac{w_1 + w_2}{\{2(1+\rho)\}^{1/2}} \geq \frac{2c}{\{2(1+\rho)\}^{1/2}}\right]$$

$$= \Pr\left[u_{12} \geq 2c/\{2(1+\rho)\}^{1/2}\right],$$

where the last equality follows from the distribution of u_{12} derived at equation (2.2.3). Hence

$$(2.5.3)\quad \Pr(w_1 > c,\ w_2 > c) \leq \frac{1}{2} I_{1-2c^2/(1+\rho)}\left(\frac{p-1}{2}, \frac{1}{2}\right).$$

Table 2.5.1, gives the values of the bound given in equation (2.5.2), viz, $[\Pr(w_i > c)]^2$ which does not depend on ρ as long as $\rho \leq 0$. Tables 2.5.2, 2.5.3 and 2.5.4 give the exact values of $M(c,c,\rho,p)$ for $c = 0.10\ (0.10)\ 0.80$, $p = 2(1)\ 10$ and $\rho = -0.5, 0$ and 0.5 respectively. Corresponding values of the bound given in equation (2.5.3) are also shown in these tables. From these tables we observe that the bound given in equation (2.5.2) is better for small values of c compared to the other one, when $\rho \leq 0$. But when c is close to the value $[(1+\rho)/2]^{1/2}$, then the other bound is better. Also for $\rho > 0$, the only bound for the bivariate probability is given by (2.5.3). Thus both bounds are useful depending upon the values of ρ, c and p. But for practical purposes, it is better to calculate the exact bivariate probabilities by using the relations given in Theorem 2.4.3 or by using numerical integration.

TABLE 2.3.1. Nominal upper critical values u_α of one-sided test statistic U for two outliers in linear regression.

(α = 0.005)

n|k	1	2	3	4	5	6	7
5	.9911	.9990	1.0000				
6	.9788	.9932	.9993	1.0000			
7	.9636	.9821	.9946	.9995	1.0000		
8	.9470	.9676	.9845	.9955	.9996	1.0000	
9	.9301	.9513	.9707	.9864	.9962	.9997	1.0000
10	.9133	.9345	.9549	.9732	.9878	.9968	.9998
11	.8970	.9177	.9382	.9578	.9753	.9890	.9972
12	.8813	.9019	.9214	.9414	.9603	.9770	.9899
14	.8519	.8701	.8890	.9083	.9276	.9466	.9644
15	.8382	.8556	.8736	.8922	.9112	.9302	.9488
16	.8251	.8417	.8589	.8768	.8952	.9139	.9326
18	.8008	.8158	.8315	.8479	.8648	.8824	.9003
20	.7786	.7923	.8066	.8215	.8370	.8532	.8699
21	.7682	.7813	.7949	.8092	.8240	.8395	.8555
24	.7396	.7511	.7631	.7755	.7885	.8020	.8161
25	.7309	.7419	.7533	.7653	.7777	.7907	.8042
27	.7143	.7245	.7351	.7460	.7575	.7694	.7817
28	.7065	.7163	.7264	.7370	.7480	.7594	.7713
30	.6917	.7008	.7102	.7200	.7301	.7406	.7516
32	.6779	.6864	.6951	.7042	.7136	.7233	.7335
33	.6714	.6795	.6880	.6967	.7058	.7152	.7249
35	.6589	.6665	.6744	.6826	.6910	.6998	.7088
36	.6529	.6603	.6680	.6758	.6840	.6925	.7012
40	.6308	.6373	.6441	.6510	.6582	.6656	.6733
42	.6206	.6268	.6332	.6397	.6465	.6534	.6606
45	.6064	.6121	.6179	.6240	.6301	.6365	.6431
48	.5932	.5985	.6039	.6094	.6151	.6210	.6270
49	.5890	.5942	.5994	.6048	.6104	.6161	.6220
50	.5849	.5899	.5951	.6004	.6058	.6113	.6170
55	.5658	.5703	.5749	.5795	.5843	.5892	.5943
60	.5487	.5527	.5568	.5610	.5653	.5697	.5741
70	.5191	.5224	.5258	.5292	.5327	.5363	.6399
80	.4943	.4971	.4999	.5028	.5057	.5087	.5117
90	.4731	.4755	.4779	.4803	.4828	.4854	.4880
100	.4546	.4567	.4588	.4610	.4631	.4653	.4676

TABLE 2.3.1 Contd.

(α = 0.005)

n|k	8	9	10	11	12	13	14	15
10	1.0000							
11	.9998	1.0000						
12	.9975	.9998	1.0000					
14	.9798	.9914	.9980	.9999	1.0000			
15	.9661	.9809	.9920	.9982	.9999	1.0000		
16	.9507	.9676	.9819	.9925	.9983	.9999	1.0000	
18	.9185	.9366	.9541	.9701	.9836	.9934	.9986	.9999
20	.8871	.9047	.9225	.9400	.9569	.9722	.9850	.9941
21	.8721	.8892	.9067	.9242	.9415	.9581	.9731	.9856
24	.8308	.8460	.8618	.8781	.8948	.9118	.9288	.9455
25	.8182	.8328	.8480	.8637	.8799	.8965	.9133	.9301
27	.7946	.8081	.8220	.8365	.8516	.8671	.8831	.8995
28	.7836	.7965	.8099	.8238	.8382	.8532	.8687	.8846
30	.7630	.7748	.7871	.7999	.8133	.8271	.8415	.8563
32	.7440	.7549	.7663	.7781	.7903	.8031	.8164	.8301
33	.7351	.7456	.7565	.7678	.7796	.7919	.8046	.8178
35	.7182	.7279	.7381	.7485	.7594	.7708	.7825	.7947
36	.7103	.7196	.7294	.7395	.7500	.7608	.7721	.7839
40	.6811	.6893	.6978	.7065	.7155	.7249	.7346	.7447
42	.6680	.6757	.6836	.6917	.7002	.7089	.7180	.7273
45	.6498	.6568	.6640	.6714	.6790	.6869	.6951	.7035
48	.6332	.6396	.6462	.6529	.6599	.6671	.6745	.6822
49	.6280	.6342	.6406	.6471	.6539	.6609	.6681	.6755
50	.6229	.6289	.6351	.6415	.6481	.6549	.6618	.6690
55	.5994	.6047	.6102	.6157	.6215	.6273	.6334	.6396
60	.5787	.5834	.5882	.5931	.5982	.6034	.6087	.6141
70	.5436	.5474	.5513	.5552	.5593	.5634	.5676	.5719
80	.5148	.5180	.5212	.5244	.5278	.5311	.5346	.5381
90	.4906	.4933	.4960	.4987	.5015	.5044	.5073	.5102
100	.4698	.4721	.4745	.4768	.4793	.4817	.4842	.4867

TABLE 2.3.1 Contd.

(α = 0.01)

n|k	1	2	3	4	5	6	7
5	.9859	.9980	1.0000				
6	.9700	.9893	.9987	1.0000			
7	.9518	.9747	.9914	.9990	1.0000		
8	.9330	.9571	.9781	.9929	.9993	1.0000	
9	.9144	.9385	.9613	.9807	.9940	.9994	1.0000
10	.8964	.9198	.9430	.9646	.9827	.9948	.9996
11	.8791	.9016	.9244	.9467	.9673	.9844	.9955
12	.8628	.8841	.9061	.9283	.9499	.9697	.9858
14	.8325	.8517	.8717	.8924	.9136	.9347	.9550
15	.8186	.8367	.8557	.8755	.8959	.9167	.9374
16	.8054	.8226	.8406	.8594	.8789	.8991	.9195
18	.7809	.7964	.8126	.8295	.8473	.8658	.8849
20	.7588	.7727	.7873	.8027	.8187	.8355	.8530
21	.7484	.7617	.7757	.7902	.8055	.8215	.8382
24	.7202	.7317	.7438	.7565	.7697	.7835	.7979
25	.7115	.7226	.7342	.7462	.7589	.7720	.7858
27	.6952	.7054	.7160	.7271	.7386	.7507	.7633
28	.6875	.6973	.7075	.7181	.7292	.7407	.7528
30	.6730	.6821	.6915	.7013	.7115	.7221	.7331
32	.6595	.6679	.6767	.6857	.6952	.7049	.7151
33	.6531	.6612	.6697	.6784	.6875	.6969	.7067
35	.6409	.6485	.6563	.6645	.6729	.6816	.6907
36	.6351	.6424	.6500	.6579	.6660	.6744	.6832
40	.6135	.6200	.6266	.6336	.6407	.6481	.6557
42	.6036	.6097	.6160	.6225	.6292	.6361	.6433
45	.5897	.5953	.6011	.6071	.6132	.6195	.6261
48	.5769	.5821	.5874	.5929	.5986	.6044	.6103
49	.5728	.5779	.5831	.5884	.5939	.5996	.6054
50	.5688	.5738	.5789	.5841	.5894	.5949	.6006
55	.5503	.5547	.5592	.5638	.5685	.5734	.5784
60	.5337	.5376	.5416	.5458	.5500	.5543	.5587
70	.5049	.5082	.5115	.5149	.5183	.5218	.5254
80	.4809	.4836	.4864	.4892	.4921	.4950	.4980
90	.4603	.4627	.4651	.4675	.4699	.4724	.4750
100	.4425	.4445	.4466	.4487	.4508	.4530	.4552

TABLE 2.3.1. Contd.

(α = 0.01)

n|k	8	9	10	11	12	13	14	15
10	1.0000							
11	.9996	1.0000						
12	.9960	.9997	1.0000					
14	.9733	.9879	.9968	.9998	1.0000			
15	.9571	.9748	.9887	.9971	.9998	1.0000		
16	.9397	.9590	.9761	.9894	.9973	.9998	1.0000	
18	.9045	.9243	.9438	.9622	.9784	.9907	.9977	.9999
20	.8712	.8900	.9092	.9284	.9473	.9649	.9802	.9916
21	.8555	.8736	.8922	.9112	.9302	.9488	.9661	.9809
24	.8131	.8288	.8453	.8624	.8800	.8982	.9166	.9349
25	.8002	.8153	.8310	.8474	.8644	.8819	.8999	.9182
27	.7764	.7901	.8045	.8194	.8350	.8512	.8681	.8854
28	.7653	.7784	.7921	.8064	.8213	.8369	.8530	.8698
30	.7446	.7566	.7692	.7822	.7959	.8101	.8249	.8403
32	.7257	.7367	.7482	.7602	.7727	.7857	.7993	.8134
33	.7163	.7274	.7384	.7499	.7619	.7744	.7874	.8009
35	.7001	.7099	.7201	.7307	.7417	.7531	.7651	.7775
36	.6923	.7017	.7115	.7216	.7322	.7432	.7546	.7665
40	.6635	.6717	.6801	.6889	.6979	.7074	.7171	.7273
42	.6506	.6582	.6661	.6743	.6827	.6915	.7005	.7099
45	.6328	.6397	.6468	.6542	.6618	.6697	.6779	.6863
48	.6165	.6228	.6293	.6361	.6430	.6501	.6575	.6652
49	.6114	.6175	.6239	.6304	.6371	.6440	.6512	.6586
50	.6064	.6124	.6185	.6249	.6314	.6381	.6451	.6522
55	.5835	.5887	.5941	.5996	.6052	.6111	.6171	.6232
60	.5632	.5679	.5726	.5775	.5825	.5876	.5928	.5982
70	.5291	.5328	.5366	.5405	.5445	.5485	.5527	.5569
80	.5011	.5042	.5073	.5105	.5138	.5171	.5205	.5240
90	.4775	.4802	.4828	.4855	.4883	.4911	.4939	.4968
100	.4574	.4597	.4619	.4643	.4666	.4690	.4715	.4739

TABLE 2.3.1. Contd.

(α = 0.025)

n \ k	1	2	3	4	5	6	7
5	.9740	.9950	1.000				
6	.9525	.9802	.9967	1.0000			
7	.9302	.9599	.9842	.9976	1.0000		
8	.9085	.9379	.9653	.9870	.9982	1.0000	
9	.8879	.9160	.9439	.9694	.9890	.9986	1.0000
10	.8686	.8951	.9222	.9487	.9727	.9905	.9989
11	.8504	.8752	.9011	.8273	.9527	.9753	.9917
12	.8335	.8567	.8810	.9062	.9317	.9561	.9774
14	.8026	.8230	.8444	.8670	.8905	.9147	.9387
15	.7886	.8077	.8278	.8491	.8713	.8945	.9182
16	.7754	.7933	.8122	.8322	.8533	.8753	.8981
18	.7512	.7670	.7838	.8015	.8202	.8399	.8606
20	.7294	.7436	.7586	.7744	.7910	.8086	.8271
21	.7193	.7328	.7470	.7619	.7777	.7943	.8117
24	.6917	.7034	.7156	.7284	.7418	.7559	.7708
25	.6833	.6944	.7061	.7183	.7311	.7445	.7586
27	.6675	.6777	.6884	.6995	.7111	.7233	.7361
28	.6601	.6699	.6801	.6907	.7018	.7135	.7257
30	.6461	.6551	.6645	.6743	.6845	.6951	.7062
32	.6331	.6414	.6501	.6592	.6685	.6783	.6885
33	.6269	.6350	.6433	.6520	.6611	.6705	.6802
35	.6152	.6227	.6305	.6385	.6469	.6556	.6646
36	.6096	.6168	.6244	.6321	.6402	.6486	.6573
40	.5889	.5953	.6019	.6087	.6157	.6230	.6305
42	.5794	.5854	.5916	.5980	.6046	.6115	.6185
45	.5661	.5717	.5773	.5832	.5892	.5955	.6019
48	.5539	.5590	.5642	.5696	.5751	.5808	.5867
49	.5500	.5549	.5600	.5653	.5707	.5762	.5820
50	.5462	.5510	.5560	.5611	.5664	.5718	.5773
55	.5285	.5328	.5372	.5417	.5463	.5511	.5559
60	.5126	.5164	.5204	.5244	.5285	.5328	.5371
70	.4852	.4884	.4916	.4949	.4982	.5017	.5052
80	.4623	.4649	.4677	.4704	.4732	.4761	.4790
90	.4427	.4450	.4473	.4496	.4520	.4545	.4569
100	.4257	.4276	.4297	.4317	.4338	.4359	.4380

TABLE 2.3.1. Contd.

(α = 0.025)

n|k	8	9	10	11	12	13	14	15
10	1.0000							
11	.9991	1.0000						
12	.9926	.9992	1.0000					
14	.9614	.9808	.9941	.9995	1.0000			
15	.9416	.9636	.9821	.9946	.9995	1.0000		
16	.9213	.9442	.9655	.9833	.9951	.9996	1.0000	
18	.8821	.9043	.9267	.9486	.9687	.9852	.9958	.9997
20	.8465	.8668	.8879	.9095	.9313	.9522	.9713	.9867
21	.8301	.8494	.8695	.8904	.9118	.9332	.9538	.9725
24	.7864	.8028	.8201	.8382	.8571	.8768	.8971	.9177
25	.7734	.7890	.8054	.8226	.8406	.8594	.8789	.8991
27	.7495	.7635	.7783	.7938	.8100	.8271	.8449	.8635
28	.7384	.7518	.7658	.7805	.7960	.8122	.8292	.8470
30	.7179	.7300	.7428	.7561	.7701	.7848	.8001	.8162
32	.6992	.7103	.7219	.7340	.7467	.7601	.7740	.7886
33	.6904	.7011	.7122	.7238	.7359	.7486	.7619	.7758
35	.6741	.6838	.6940	.7047	.7158	.7274	.7395	.7522
36	.6664	.6758	.6855	.6957	.7064	.7175	.7291	.7412
40	.6384	.6465	.6549	.6636	.6726	.6820	.6918	.7020
42	.6258	.6334	.6412	.6493	.6577	.6664	.6755	.6849
45	.6085	.6154	.6224	.6297	.6373	.6451	.6532	.6617
48	.5928	.5990	.6055	.6121	.6189	.6260	.6334	.6409
49	.5878	.5939	.6001	.6066	.6132	.6201	.6272	.6345
50	.5830	.5889	.5950	.6012	.6077	.6143	.6212	.6283
55	.5609	.5661	.5714	.5768	.5824	.5881	.5940	.6001
60	.5415	.5460	.5507	.5555	.5604	.5654	.5705	.5758
70	.5087	.5124	.5161	.5199	.5238	.5278	.5318	.5360
80	.4819	.4849	.4880	.4911	.4943	.4976	.5009	.5043
90	.4594	.4620	.4646	.4672	.4699	.4726	.4754	.4782
100	.4402	.4424	.4446	.4469	.4491	.4515	.4539	.4563

TABLE 2.3.1. Contd.

(α = 0.05)

n|k	1	2	3	4	5	6	7
5	.9587	.9900	.9999				
6	.9326	.9685	.9933	.9999			
7	.9074	.9431	.9749	.9952	1.0000		
8	.8840	.9177	.9508	.9793	.9964	1.0000	
9	.8623	.8936	.9257	.9567	.9825	.9972	1.0000
10	.8424	.8712	.9014	.9321	.9613	.9849	.9978
11	.8239	.8505	.8786	.9079	.9374	.9650	.9868
12	.8069	.8314	.8575	.8850	.9135	.9419	.9680
14	.7762	.7972	.8197	.8436	.8689	.8954	.9225
15	.7623	.7819	.8029	.8251	.8488	.8737	.8997
16	.7493	.7677	.7872	.8080	.8300	.8534	.8780
18	.7256	.7417	.7588	.7770	.7964	.8169	.8386
20	.7044	.7187	.7339	.7500	.7670	.7850	.8042
21	.6946	.7081	.7225	.7376	.7537	.7706	.7886
24	.6679	.6795	.6917	.7046	.7182	.7324	.7475
25	.6597	.6708	.6825	.6947	.7076	.7211	.7354
27	.6445	.6547	.6653	.6764	.6880	.7003	.7131
28	.6374	.6471	.6572	.6678	.6789	.6906	.7028
30	.6239	.6328	.6421	.6518	.6620	.6726	.6838
32	.6113	.6196	.6282	.6372	.6465	.6562	.6664
33	.6054	.6134	.6216	.6303	.6392	.6486	.6583
35	.5941	.6015	.6092	.6172	.6255	.6341	.6431
36	.5887	.5959	.6033	.6110	.6190	.6273	.6360
40	.5688	.5751	.5816	.5884	.5953	.6025	.6100
42	.5597	.5657	.5718	.5781	.5846	.5913	.5983
45	.5470	.5524	.5580	.5638	.5697	.5759	.5822
48	.5352	.5403	.5454	.5507	.5562	.5618	.5676
49	.5315	.5364	.5414	.5466	.5519	.5573	.5629
50	.5279	.5326	.5375	.5425	.5477	.5530	.5585
55	.5109	.5151	.5194	.5239	.5284	.5331	.5378
60	.4957	.4994	.5033	.5073	.5113	.5154	.5197
70	.4694	.4725	.4757	.4789	.4822	.4855	.4889
80	.4474	.4500	.4527	.4554	.4581	.4609	.4637
90	.4287	.4309	.4331	.4354	.4378	.4401	.4426
100	.4123	.4143	.4162	.4182	.4202	.4223	.4244

TABLE 2.3.1. Contd.

(α = 0.05)

n|k	8	9	10	11	12	13	14	15
10	1.0000							
11	.9982	1.0000						
12	.9883	.9985	1.0000					
14	.9490	.9728	.9906	.9989	1.0000			
15	.9261	.9518	.9747	.9914	.9990	1.0000		
16	.9036	.9294	.9544	.9763	.9922	.9992	1.0000	
18	.8615	.8855	.9102	.9350	.9586	.9791	.9933	.9993
20	.8244	.8458	.8683	.8917	.9157	.9396	.9621	.9812
21	.8077	.8278	.8491	.8713	.8945	.9182	.9416	.9636
24	.7635	.7803	.7981	.8169	.8367	.8575	.8792	.9017
25	.7505	.7664	.7831	.8009	.8196	.8393	.8600	.8816
27	.7266	.7409	.7559	.7717	.7884	.8060	.8246	.8441
28	.7157	.7292	.7434	.7584	.7742	.7909	.8084	.8269
30	.6954	.7076	.7205	.7340	.7481	.7631	.7788	.7954
32	.6770	.6882	.6998	.7120	.7248	.7383	.7525	.7674
33	.6685	.6791	.6902	.7019	.7141	.7269	.7404	.7545
35	.6525	.6622	.6724	.6830	.6942	.7058	.7180	.7308
36	.6449	.6543	.6641	.6743	.6849	.6960	.7077	.7199
40	.6177	.6257	.6341	.6427	.6517	.6611	.6709	.6811
42	.6055	.6130	.6207	.6288	.6371	.6458	.6548	.6642
45	.5887	.5955	.6025	.6097	.6172	.6250	.6330	.6414
48	.5735	.5797	.5860	.5926	.5993	.6063	.6136	.6211
49	.5687	.5747	.5809	.5872	.5938	.6006	.6076	.6148
50	.5641	.5699	.5759	.5820	.5884	.5950	.6017	.6088
55	.5428	.5478	.5530	.5583	.5638	.5695	.5753	.5813
60	.5240	.5285	.5330	.5377	.5425	.5475	.5525	.5578
70	.4924	.4960	.4996	.5034	.5072	.5111	.5151	.5191
80	.4666	.4696	.4726	.4756	.4788	.4819	.4852	.4885
90	.4450	.4475	.4500	.4526	.4552	.4579	.4606	.4634
100	.4265	.4286	.4308	.4330	.4353	.4375	.4398	.4422

TABLE 2.3.1. Contd.

(α = 0.10)

n|k	1	2	3	4	5	6	7
5	.9343	.9800	.9995				
6	.9042	.9500	.9867	.9998			
7	.8770	.9192	.9601	.9905	.9999		
8	.8526	.8907	.9302	.9671	.9929	.9999	
9	.8306	.8649	.9014	.9385	.9722	.9944	1.0000
10	.8106	.8416	.8749	.9100	.9451	.9760	.9956
11	.7924	.8205	.8508	.8833	.9171	.9504	.9790
12	.7756	.8012	.8289	.8587	.8904	.9230	.9547
14	.7457	.7674	.7907	.8159	.8428	.8716	.9018
15	.7323	.7524	.7739	.7971	.8220	.8486	.8769
16	.7198	.7384	.7584	.7798	.8029	.8275	.8539
18	.6970	.7132	.7306	.7491	.7690	.7902	.8129
20	.6766	.6910	.7063	.7225	.7398	.7582	.7779
21	.6672	.6808	.6952	.7105	.7267	.7439	.7623
24	.6417	.6533	.6655	.6784	.6919	.7063	.7215
25	.6340	.6450	.6566	.6688	.6817	.6952	.7096
27	.6195	.6295	.6400	.6511	.6627	.6749	.6878
28	.6126	.6222	.6323	.6428	.6539	.6655	.6777
30	.5998	.6086	.6178	.6275	.6375	.6481	.6592
32	.5879	.5960	.6045	.6133	.6226	.6322	.6423
33	.5822	.5900	.5982	.6067	.6156	.6248	.6345
35	.5715	.5788	.5863	.5942	.6024	.6109	.6198
36	.5664	.5734	.5807	.5883	.5962	.6044	.6129
40	.5474	.5536	.5600	.5666	.5734	.5805	.5878
42	.5388	.5446	.5506	.5568	.5632	.5698	.5766
45	.5267	.5320	.5375	.5431	.5489	.5550	.5612
48	.5155	.5204	.5254	.5306	.5359	.5414	.5471
49	.5119	.5167	.5216	.5266	.5318	.5372	.5427
50	.5085	.5131	.5179	.5228	.5279	.5331	.5384
55	.4923	.4964	.5006	.5050	.5094	.5140	.5186
60	.4778	.4815	.4853	.4891	.4931	.4971	.5013
70	.4528	.4558	.4589	.4621	.4653	.4685	.4719
80	.4319	.4344	.4370	.4396	.4423	.4450	.4478
90	.4140	.4161	.4183	.4206	.4228	.4252	.4275
100	.3984	.4003	.4022	.4041	.4061	.4081	.4101

TABLE 2.3.1. Contd.

(α = 0.10)

n|k	8	9	10	11	12	13	14	15
10	1.0000							
11	.9964	1.0000						
12	.9814	.9970	1.0000					
14	.9324	.9615	.9850	.9978	1.0000			
15	.9065	.9362	.9642	.9864	.9981	1.0000		
16	.8817	.9106	.9396	.9665	.9875	.9983	1.0000	
18	.8371	.8629	.8899	.9177	.9453	.9703	.9894	.9987
20	.7989	.8214	.8452	.8704	.8967	.9236	.9499	.9734
21	.7819	.8029	.8251	.8488	.8737	.8997	.9261	.9518
24	.7377	.7549	.7731	.7925	.8131	.8350	.8581	.8824
25	.7248	.7410	.7581	.7763	.7956	.8162	.8379	.8609
27	.7013	.7157	.7309	.7470	.7640	.7821	.8013	.8217
28	.6906	.7042	.7185	.7337	.7498	.7668	.7848	.8040
30	.6708	.6830	.6959	.7095	.7238	.7389	.7549	.7719
32	.6529	.6640	.6756	.6879	.7007	.7143	.7286	.7437
33	.6446	.6552	.6663	.6779	.6901	.7030	.7166	.7309
35	.6291	.6387	.6489	.6594	.6705	.6822	.6944	.7073
36	.6218	.6311	.6407	.6509	.6615	.6726	.6842	.6965
40	.5955	.6034	.6116	.6202	.6291	.6384	.6481	.6583
42	.5837	.5911	.5987	.6067	.6149	.6235	.6325	.6418
45	.5676	.5742	.5811	.5883	.5956	.6033	.6113	.6195
48	.5530	.5590	.5652	.5717	.5784	.5853	.5924	.5998
49	.5484	.5542	.5603	.5665	.5730	.5797	.5866	.5937
50	.5439	.5496	.5555	.5615	.5678	.5743	.5809	.5879
55	.5235	.5284	.5335	.5387	.5441	.5497	.5554	.5613
60	.5055	.5099	.5143	.5189	.5236	.5285	.5334	.5385
70	.4753	.4787	.4823	.4859	.4897	.4935	.4974	.5014
80	.4506	.4535	.4564	.4594	.4624	.4655	.4687	.4719
90	.4299	.4323	.4348	.4373	.4398	.4424	.4451	.4478
100	.4122	.4143	.4164	.4185	.4207	.4229	.4252	.4275

TABLE 2.3.2. Nominal upper critical values u_α, for $\alpha = 0.02625$ and selected values of n and k.

n	k	u_α
10	1	0.8669
11	1	0.8487
20	8	0.8450
21	1	0.7176
48	13	0.6247

TABLE 2.3.3. Nominal upper critical values v_α, for $\alpha = 0.02625$ and selected values of n and k.

n	k	v_α
20	1	0.7504
20	8	0.8643
48	13	0.6432

TABLE 2.5.1. Bound given at equation (2.5.2) for $M(c,c,\rho,p)$ when $\rho \leq 0$.

p|c	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80
2	0.21913	0.19001	0.16242	0.13617	0.11111	0.08712	0.06410	0.04196
3	0.20250	0.16000	0.12250	0.09000	0.06250	0.04000	0.02250	0.01000
4	0.19048	0.13952	0.09729	0.06366	0.03822	0.02027	0.00885	0.00271
5	0.18084	0.12390	0.07938	0.04666	0.02441	0.01082	0.00369	0.00078
6	0.17267	0.11129	0.06587	0.03496	0.16024	0.00596	0.00160	0.02364
7	0.16555	0.11008	0.05530	0.02660	0.01072	0.00335	0.00071	0.00007
8	0.15919	0.09179	0.04684	0.02046	0.00727	0.00192	0.00032	0.00002
9	0.15344	0.08398	0.03994	0.01589	0.00498	0.00111	0.00015	0.00001
10	0.14818	0.08398	0.03424	0.01242	0.00344	0.00065	0.00007	0.00000

TABLE 2.5.2. Exact value of $M(c,c,\rho,p)$ in top row and its bound given at equation (2.5.3) in bottom row for $\rho = -0.5$.

p | c	0.10	0.20	0.30	0.40	0.50
2	0.13478	0.10257	0.06968	0.03568	0.00000
	0.43591	0.36901	0.29517	0.20483	0.00000
3	0.11944	0.07793	0.04274	0.01527	0.00000
	0.40000	0.30000	0.20000	0.10000	0.00000
4	0.10862	0.06225	0.02821	0.00720	0.00000
	0.37353	0.25232	0.14238	0.05204	0.00000
5	0.10015	0.05107	0.01938	0.00357	0.00000
	0.35200	0.21600	0.10400	0.02800	0.00000
6	0.09312	0.04262	0.01363	0.00183	0.00000
	0.33361	0.18697	0.07719	0.01537	0.00000
7	0.08712	0.03599	0.00976	0.00096	0.00000
	0.31744	0.16308	0.05792	0.00856	0.00000
8	0.08186	0.03065	0.00708	0.00051	0.00000
	0.30295	0.14305	0.04381	0.00481	0.00000
9	0.07720	0.02629	0.00519	0.00028	0.00000
	0.28979	0.12604	0.03334	0.00273	0.00000
10	0.07301	0.02268	0.00383	0.00015	0.00000
	0.27772	0.11143	0.02550	0.00156	0.00000

TABLE 2.5.3. Exact value of $M(c,c,\rho,p)$ in top row and its bound given at equation (2.5.3) in bottom row for $\rho = 0$.

p|c	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80
2	0.21812 0.45483	0.18591 0.40872	0.15301 0.36053	0.11901 0.30861	0.08333 0.25000	0.04517 0.17749	0.00318 0.04517	0.00000 0.00000
3	0.20160 0.42929	0.15645 0.35858	0.11478 0.28787	0.07700 0.21716	0.04387 0.14645	0.01689 0.07574	0.00030 0.00503	0.00000 0.00000
4	0.18963 0.41027	0.13626 0.32237	0.09057 0.23824	0.05325 0.16013	0.02508 0.09085	0.00697 0.03457	0.00003 0.00060	0.00000 0.00000
5	0.18001 0.39464	0.12084 0.29352	0.07340 0.20089	0.03815 0.12099	0.01499 0.05806	0.00304 0.01634	0.00000 0.00008	0.00000 0.00000
6	0.17186 0.38115	0.10840 0.26940	0.06051 0.17139	0.02796 0.09282	0.00921 0.03779	0.00137 0.00789	0.00000 0.00001	0.00000 0.00000
7	0.16475 0.36917	0.09801 0.24864	0.05048 0.14740	0.02082 0.07195	0.00578 0.02491	0.00063 0.00387	0.00000 0.00000	0.00000 0.00000
8	0.15841 0.35833	0.08915 0.23042	0.04249 0.12753	0.01568 0.05620	0.00367 0.01657	0.00030 0.00191	0.00000 0.00000	0.00000 0.00000
9	0.15267 0.34838	0.08147 0.21422	0.03600 0.11085	0.01192 0.04415	0.00236 0.01110	0.00014 0.00096	0.00000 0.00000	0.00000 0.00000
10	0.14742 0.33915	0.07473 0.19968	0.03068 0.09671	0.00913 0.03485	0.00153 0.00748	0.00007 0.00048	0.00000 0.00000	0.00000 0.00000

TABLE 2.5.4. Exact value of M(c,c,ρ,p) in top row and its bound given at equation (2.5.3) in bottom row for ρ = 0.5.

p|c	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80
2	0.30145 0.46316	0.26924 0.42582	0.23635 0.38740	0.20234 0.34718	0.16667 0.30409	0.12850 0.25637	0.08652 0.20039	0.03817 0.12510
3	0.28425 0.44227	0.23705 0.38453	0.19182 0.32679	0.14876 0.26906	0.10817 0.21132	0.07063 0.15359	0.03716 0.09585	0.01014 0.03812
4	0.27162 0.42665	0.21421 0.35430	0.16179 0.28396	0.11505 0.21678	0.07478 0.15403	0.04187 0.09734	0.01744 0.04891	0.00299 0.01249
5	0.26133 0.41378	0.19624 0.32987	0.13926 0.25058	0.09142 0.17822	0.05346 0.11510	0.02585 0.06352	0.00859 0.02581	0.00094 0.00425
6	0.25253 0.40264	0.18132 0.30916	0.12141 0.22328	0.07388 0.14839	0.03904 0.08734	0.01638 0.04221	0.00437 0.01390	0.00030 0.00148
7	0.24476 0.39271	0.16854 0.29107	0.10680 0.20029	0.06042 0.12463	0.02894 0.06699	0.01056 0.02840	0.00226 0.00759	0.00010 0.00052
8	0.23777 0.38368	0.15737 0.27497	0.09458 0.18056	0.04985 0.10535	0.02169 0.05178	0.00690 0.01928	0.00119 0.00419	0.00003 0.00019
9	0.23139 0.37538	0.14746 0.26046	0.08419 0.16340	0.04142 0.08950	0.01639 0.04026	0.00456 0.01318	0.00064 0.00233	0.00001 0.00007
10	0.22550 0.36765	0.13856 0.24723	0.07526 0.14833	0.03460 0.07633	0.01247 0.03145	0.00303 0.00905	0.00034 0.00130	0.00000 0.00002

CHAPTER III

TWO OUTLIERS IN A RANDOM SAMPLE FROM $N(\mu,\sigma^2)$

3.1. Introduction

Let $y_1, y_2, \ldots, y_n$ be n independently and normally distributed observations which constitute a random sample of size n from a $N(\mu,\sigma^2)$ distribution. In this chapter, we shall denote the ith order statistic by $y_{(i)}$, where $y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(n)}$ are obtained by rearranging $y_1, y_2, \ldots, y_n$ in an increasing order of magnitude. Now, the residual vector is $\underset{\sim}{e} = (e_1, e_2, \ldots, e_n)'$, where

$$e_i = y_i - \bar{y}$$

$$\bar{y} = \sum_{i=1}^{n} y_i/n$$

is the sample mean, and the error sum of squares $S^2 = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i-\bar{y})^2$ is based on n-1 degrees of freedom.

The elements of variance covariance matrix $\underset{\sim}{\Lambda}$ are given by

$$(3.1.1) \qquad \lambda_{ij} = \begin{cases} -1/n & \text{if } i \neq j \\ (n-1)/n & \text{if } i = j, \end{cases}$$

and the correlation coefficient is $\rho_{ij} = -1/(n-1) = \rho$.
Denote the common variance of residuals by $\lambda = (n-1)/n$. Hence the studentized residual as defined in (1.2.5) is given by

$$(3.1.2) \qquad w_i = e_i/(S_p \lambda^{1/2}), \quad i = 1,2,\ldots,n,$$

where $S_p^2 = S^2 + \nu s_\nu^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 + \nu s_\nu^2$ is the pooled sum of squares based on $p = n-1+\nu$ degrees of freedom.

Equation (2.1.1) now reduces to

$$u_{ij} = (e_i+e_j)/[S_p \{2\lambda(1+\rho)\}^{1/2}]$$

$$= [n/\{2(n-2)\}]^{1/2} [(y_i+y_j-2\bar{y})/S_p] .$$

This gives

$$U = \max_{1 \le i < j \le n} u_{ij} = [n/\{2(n-2)\}]^{1/2} [(y_{(n)}+y_{(n-1)}-2\bar{y})/S_p]$$

$$= [n/\{2(n-2)\}]^{1/2} M,$$

where $M = (y_{(n)}+y_{(n-1)}-2\bar{y})/S_p$ is the Murphy's statistic for two outliers. This has been studied in detail by Hawkins (1978). He also provides exact upper 100α percent of M for n = 5(1)15(5)30, α = 0.001, 0.01, 0.05 and 0.1; and ν = 0, 5, 15 and 30. Power studies for M has also been done by him. Note that for ν = 0, Murphy's statistic is commonly defined by (Barnett and Lewis, 1978)

$$T_{N3} = (y_{(n)} + y_{(n-1)} - 2\bar{y})/s,$$

where $s^2 = S^2/(n-1)$. Consequently for ν = 0,

(3.1.3) $$U = [n/\{2(n-1)(n-2)\}]^{1/2} \cdot T_{N3}.$$

Simulated upper percentage points of T_{N3} are tabulated by Barnett and Lewis for n = 5(1)10(2) 20(10) 50, 100, and α = 0.01 and 0.05.

Similarly, for the two sided statistic V, we have

$$v_{ij} = (y_i - y_j)/(S_p\, 2^{1/2}), \text{ and}$$

$$V = \max_{1 \le i < j \le n} |v_{ij}| = (y_{(n)} - y_{(1)})/(S_p\, 2^{1/2})$$

$$= (y_{(n)} - y_{(1)})/[2\{(n-1)s^2 + \nu s_\nu^2\}]^{1/2} .$$

For $\nu = 0$, V reduces to

$$(3.1.4) \qquad V = [1/\{2(n-1)\}^{1/2}] \cdot T_{N6} ,$$

where $T_{N6} = (y_{(n)} - y_{(1)})/s$ is the internally studentized range statistic.

This statistic T_{N6} has been used as the discordancy test for a lower and upper outlier-pair $y_{(1)}, y_{(n)}$ in a normal sample with μ and σ^2 unknown, by David, Hartley and Pearson (1954), Pearson and Stephens (1964), Shapiro, Wilk and Chen (1968), Barnett and Lewis (1978) etc.

The above calculations show that our statistics U and V reduce to the widely used statistics M and T_{N6} respectively for $\nu = 0$.

3.2. Distribution theory

In this case the marginal and joint distributions are immediately obtained from the general results derived in Section 2.2, with k = 1 and $\rho_{ij} = \rho = -1/(n-1)$ for all $i \neq j$.

The marginal distribution of u_{ij} is given at (2.2.11) with $p = n-1+\nu$. The joint density of u_{ij} and $u_{i_1 j_1}$ as given

at (2.2.9) reduces to

$$(3.2.1)\quad f(u_{ij},u_{i_1j_1}) = \frac{(n-3+\nu)}{2\pi(1-\rho_1^2)^{1/2}}\left[1 - \frac{1}{1-\rho_1^2}\left(u_{ij}^2 - 2\rho_1 u_{ij}\, u_{i_1j_1} + u_{i_1j_1}^2\right)\right]^{(n-5+\nu)/2}$$

and the function is defined over

$$u_{ij}^2 - 2\rho_1\, u_{ij}\, u_{i_1j_1} + u_{i_1j_1}^2 \leq 1 - \rho_1^2 ,$$

where the shape parameter from equation (2.2.8) is

$$\rho_1 = \frac{\rho_{ii_1} + \rho_{ij_1} + \rho_{i_1j} + \rho_{jj_1}}{2\left[(1+\rho_{ij})(1+\rho_{i_1j_1})\right]^{1/2}} .$$

In particular when $i = 1$, $j = 2$, $i_1 = 1$, $j_1 = 3$, that is for the joint distribution of u_{12} and u_{13}, we have

$$(3.2.2)\quad \rho_1 = (1+3\rho)/[2(1+\rho)] = (n-4)/[2(n-2)]$$

Similarly, for the joint distribution of (u_{12},u_{34}) we have

$$(3.2.3)\quad \rho_1 = 4\rho/[2(1+\rho)] = -2/(n-2) .$$

Let $N = n(n-1)/2$, which is the total number of distinct u_{ij}'s used in determing the statistic U. The total number of distinct combinations of two u_{ij}'s are given by

$$(3.2.4)\quad \binom{N}{2} = \frac{N(N-1)}{2} = \frac{n(n-1)(n-2)(n+1)}{8} .$$

Out of these pairs, for u_{ij} and $u_{i_1j_1}$, there are 4 different types of combinations, which are given below :

(i) $i = i_1,\ j \neq j_1$, like u_{12}, u_{13}

(ii) $i \neq i_1,\ j = j_1$, like u_{13}, u_{23}

(iii) $i \neq i_1,\ j = i_1,\ j \neq j_1$, like u_{12}, u_{23}

(iv) $i \neq i_1 \neq j_1 \neq j_2$, like u_{12}, u_{34}.

These types along with their shape parameter values and the number of combinations are given in Table 3.2.1.

TABLE 3.2.1. Types of combinations of two u_{ij}'s with corresponding shape parameter values.

Type	No. of combinations	Shape parameter
u_{12}, u_{13}	$n(n-1)(n-2)/6$	$(n-4)/[2(n-2)]$
u_{13}, u_{23}	$n(n-1)(n-2)/6$	$(n-4)/[2(n-2)]$
u_{12}, u_{23}	$n(n-1)(n-2)/6$	$(n-4)/[2(n-2)]$
u_{12}, u_{34}	$n(n-1)(n-2)(n-3)/8$	$-2/(n-2)$

From Table 3.2.1 it is clear that essentially there are only 2 types of pairs for u_{ij}'s having shape parameter values $(n-4)/[2(n-2)]$ and $-2/(n-2)$. In fact three types of pairs give same values of shape parameter and we can use condensed Table 3.2.2. Such condensed tables are useful for the general case, where it is not possible to count the number of pairs directly.

Similarly, we can obtain shape parameters for v_{ij}'s also. These are given in Table 3.2.3. The condensed table is given in Table 3.2.4.

TABLE 3.2.2. Condensed table for the combinations of two u_{ij}'s with corresponding shape parameter values.

Serial No. i	Shape parameter ρ_i	Frequency
1	$\rho_1 = (n-4)/[2(n-2)]$	$n(n-1)(n-2)/2$
2	$\rho_2 = -2/(n-2)$	$n(n-1)(n-2)(n-3)/8$

TABLE 3.2.3. Types of combinations of two v_{ij}'s with corresponding shape parameter values.

Type	No. of combinations	Shape parameter
v_{12}, v_{13}	$n(n-1)(n-2)/6$	1/2
v_{13}, v_{23}	$n(n-1)(n-2)/6$	1/2
v_{12}, v_{23}	$n(n-1)(n-2)/6$	-1/2
v_{12}, v_{34}	$n(n-1)(n-2)(n-3)/8$	0

TABLE 3.2.4. Condensed table for the combinations of two v_{ij}'s with corresponding shape parameter values.

Serial No. i	Shape parameter ρ'_i	Frequency
1	$\rho'_1 = 1/2$	$n(n-1)(n-2)/3$
2	$\rho'_2 = -1/2$	$n(n-1)(n-2)/6$
3	$\rho'_3 = 0$	$n(n-1)(n-2)(n-3)/8$

3.3. Comparison of percentile points with tabulated values

The upper and lower limits for the true percentage points can be obtained by using Bonferroni inequalities.

Nominal upper percentile points of U and V are given by the equations (2.3.1) and (2.3.2) respectively. These are given in Table 2.3.1 for $\nu = 0$ and different values of n and α.

For a lower bound of the true percentile point, we use equation (2.3.3) etc. Thus for the statistic U we need to solve the second Bonferroni inequality

$$(3.3.1) \quad \binom{n}{2} \Pr(u_{ij} > u) - \Sigma\Sigma \Pr(u_{ij} > u,\ u_{i_1 j_1} > u) = \alpha,$$

where the double sum is over all distinct terms, which appear in the second term of Bonferroni inequality. Clearly, if <u>all</u> the bivariate probabilities appearing in equation (3.3.1) are zero, then $u_\alpha = U_\alpha(e)$, the exact critical point. From equation (2.4.12) we see that a sufficient condition for the bivariate probability $M(h,h,\rho_1,p)$ to be equal to zero is that

$$\rho_1 \leq 2h^2 - 1,$$

or equivalently

$$h > [(1+\rho_1)/2]^{1/2}.$$

McMillan (1971) has shown that this condition holds for statistic M if $M_\alpha > [(3n-8)/(2n)]^{1/2}$, where $M_\alpha \equiv [2(n-2)/n]^{1/2} u_\alpha$, that is if $u_\alpha > [(3n-8)/\{4(n-2)\}]^{1/2}$. For $\nu = 0$, this is satisfied for $n \leq 10$ when $\alpha = 0.05$ and $n \leq 13$, when $\alpha = 0.01$. For slightly larger n he states that the error is negligible. We

now obtain the same result by considering the joint distributions. From Table 3.2.2, we see that the shape parameters are $\rho_1 = (n-4)/[2(n-2)]$ and $\rho_2 = -2/(n-2)$, hence, the bivariate probabilities appearing in equation (3.3.1) are zero, only if $u_\alpha > \max\ [\{(1+\rho_1)/2\}^{1/2}, \{(1+\rho_2)/2\}^{1/2}]$. Further $\rho_2 < \rho_1$, and hence we get the exact critical values if

$$u_\alpha > [(1+\rho_1)/2]^{1/2}.$$

Substituting for ρ_1, we get the condition as

$$u_\alpha > [(3n-8)/\{4(n-2)\}]^{1/2} = a_n \text{ (say)}.$$

Table 3.3.1 gives the values of a_n for n = 5(1) 15.

TABLE 3.3.1. Value of $a_n = [(3n-8)/\{4(n-2)\}]^{1/2}$ for determining the exact critical values.

n	a_n
5	0.76376
6	0.79057
7	0.80623
8	0.81650
9	0.82375
10	0.82916
11	0.83333
12	0.83666
13	0.83937
14	0.84163
15	0.84353

The nominal critical points u_α for n, α, $k = 1$ and $\nu = 0$ are already tabulated in Table 2.3.1. However, for comparison purposes, we again tabulate them for $\alpha = 0.01, 0.05$; $\nu = 0, 5$; $k = 1$, and $n = 5(1)\ 15\ (5)\ 60$ in Table 3.3.2. Comparing from Table 3.3.1, we see that for $\nu = 0$ the critical point u_α is exact for n not exceeding 10, 13 and 14 when $\alpha = 0.05$, 0.01 and 0.005 respectively. Similarly for $\nu = 5$, the values are exact for n upto 5 and 7 when $\alpha = 0.05$ and 0.01 respectively.

For comparison of other values, we evaluate $M_\alpha = [2(n-2)/n]^{1/2}\ u_\alpha$. These are also given in Table 3.3.2. Exact critical values of Murphy's statistic M are tabulated by Hawkins (1978). Exact percentile points of M given by Hawkins for $\nu = 0$, $\alpha = 0.05$ and $n = 10$, 20 and 30 are 1.066, 0.944 and 0.848 respectively. From Table 3.3.2, the corresponding values of M_α obtained through u_α are 1.066 , 0.945 and 0.852 respectively. Similarly for $\nu = 5$ and $\alpha = 0.05$ the critical values given by Hawkins for $n = 10$, 20 and 30 are 0.918 (0.917), 0.860 (0.863) and 0.793 (0.798) respectively, where the values shown within brackets are taken from Table 3.3.2. In general our values agree upto two decimal places in all the cases and upto three decimal places for small values of n.

In Table 3.3.3 we have obtained the nominal critical values of T_{N3} through u_α. For comparison purposes we have also tabulated simulated values of T_{N3} as given by Barnett and Lewis (1978) for $n = 10$, 20, 50 and 100 when $\alpha = 0.01$ and 0.05.

Here again we find that our values closely agree with the simulated values. There is some deviation for large values of n, which is due to the fact that our values are evaluated with the assumption of zero bivariate probabilities. This is not true when n is large.

Similarly the internally studentized range statistic T_{N6} is related to V as mentioned in (3.1.4). Since $|\rho_1'| = 1/2$ is the maximum of all $|\rho_i'|$'s given in Table 3.2.4, hence on applying condition given at equation (2.4.13), we see that the critical values v_α are exact if

$$v_\alpha > [(1+\rho_1')/2]^{1/2} = 3^{1/2}/2 = 0.8660..$$

Nominal critical points v_α for n, α, k = 1 and ν = 0 can be obtained from Table 2.3.1. Again for comparison purposes, we tabulate them for α = 0.01, 0.05; ν = 0, k = 1 and n = 3(1) 10(1) 20(10) 100 along with the corresponding values $[2(n-1)]^{1/2} v_\alpha$ of T_{N6} obtained through v_α in Table 3.3.4. From these tables it follows that v_α is exact for $n \leq 10$ when α = 0.05 and for $n \leq 13$ when α = 0.01.

The critical values of T_{N6} for n = 5, 16, 20 and 60 given by Barnett and Lewis (1978) with our values within brackets are 2.75(2.755), 4.24(4.247), 4.49(4.496) and 5.51(5.568) when α = 0.05 and 2.80(2.803), 4.52(4.519), 4.80(4.800) and 5.94(5.960) when α = 0.01 respectively. Here we find that nominal critical points of T_{N6} agree almost completely with the

exact critical values for small values of n, and there is a very small deviation for large n.

3.4. Lower bound for type I error probability

The number of combinations of u_{ij}'s or v_{ij}'s obtained in Section 3.2 is used for obtaining lower bounds for type I error probabilities of U and V statistics in sub-sections 3.4.1 and 3.4.2 respectively.

3.4.1. Lower bound for type I error probability of U

Using the notations introduced in Section 2.4, we see that the probability

$$Pr(u_{ij} > h,\ u_{i_1 j_1} > h | \rho_1 = \rho, p) = M(h,h,\rho,p).$$

For notational convenience denote the u_{ij}'s with single subscript u_i's $(i = 1,2,\ldots,N)$, where $N = n(n-1)/2$.

Using Bonferroni inequalities, we have

$$(3.4.1) \qquad S_1 - S_2 \leq Pr(U > u_\alpha) \leq S_1,$$

where $S_1 = \sum_{i=1}^{N} Pr(u_i > u_\alpha)$ and $S_2 = \sum\sum_{1 \leq i < j \leq N} Pr(u_i > u_\alpha,\ u_j > u_\alpha)$.

Note that $S_1 = \alpha$, while from Table 3.2.2, we have

$$S_2 = [n(n-1)(n-2)/2]\ Pr(u_i > u_\alpha,\ u_j > u_\alpha | \rho_1, p)$$
$$+ [n(n-1)(n-2)(n-3)/8]\ Pr(u_i > u_\alpha,\ u_j > u_\alpha | \rho_2, p),$$

where ρ_1 and ρ_2 are as given in Table 3.2.2.

Hence

$$S_2 = [n(n-1)(n-2)/8]\ [4M(u_\alpha, u_\alpha, (n-4)/\{2(n-2)\}, p) + (n-3)\ M(u_\alpha, u_\alpha, -2/(n-2), p)]\ .$$

The bivariate probabilities can be calculated using the method described in Section 2.4. Using this value of S_2 in the expression (3.4.1), we get a lower bound for $Pr(U > u_\alpha)$. These lower bounds for the type I error probability of U are shown in Table 3.4.1 for $\alpha = 0.01, 0.05$; $\nu = 0, 5$; and $n = 5(1)\ 15(5)\ 50$. This table also shows that nominal upper percentage points are quite close to the true upper points for $n \leq 30$. The lower limit decreases very rapidly as n increases.

Such a rapid decrease indirectly shows that equation (2.3.3) will not have a solution for large values of n. This is due to the fact that $S_1 - S_2$ appearing in equation (3.4.1) becomes negative for large n. Consequently, we do not have a systematic procedure for determining a lower limit $u_{*\alpha}$ for the exact critical value $U_\alpha(e)$ in such cases. In Section 3.5, we describe a method for obtaining approximate critical values of Murphy's test statistic. The approximation is remarkably good, and there is no need to calculate the lower limit $u_{*\alpha}$ for large n.

3.4.2. Lower bound for type I error probability of V

From equation (2.4.10) we now have

$$(3.4.2)\quad Pr(|v_{ij}| > h, |v_{i_1 j_1}| > h | \rho_1' = \rho, p) = 2\ [M(h, h, \rho, p) + M(h, h, -\rho, p)].$$

Similar to U, we denote the v_{ij}'s with single subscripts v_i's $(i = 1,2,\ldots,N)$.

The Bonferroni inequality then gives

$$(3.4.3)\quad S_1 - S_2 \leq \Pr(V > v_\alpha) \leq S_1,$$

where $S_1 = \sum_{i=1}^{N} \Pr(|v_i| > v_\alpha) = \alpha$ and

$$S_2 = \sum_{1\leq i<j\leq N} \Pr(|v_i| > v_\alpha,\ |v_j| > v_\alpha).$$

In this case, using Table 3.2.4, and equation (3.4.2), we get

$$S_2 = 2\left[n(n-1)(n-2)/2\right]\left[M(v_\alpha,v_\alpha,1/2,p) + M(v_\alpha,v_\alpha,-1/2,p) + \{(n-3)/2\}\, M(v_\alpha,v_\alpha,0,p)\right]$$

$$= n(n-1)(n-2)\left[M(v_\alpha,v_\alpha,1/2,p) + M(v_\alpha,v_\alpha,-1/2,p) + \{(n-3)/2\}\, M(v_\alpha,v_\alpha,0,p)\right].$$

We get a lower bound for required probability by substituting this value of S_2 in (3.4.3) with $S_1 = \alpha$. These lower bounds are tabulated in Table 3.4.2 for $\alpha = 0.01, 0.05$; $\nu = 0$ and $n = 10(2)\ 20(10)\ 50$. For $n \leq 10$, the lower bound is equal to α, as for these values of n and α the nominal percentage points are exact. Similar to the case of U, here also the lower limit decreases rapidly as n increases.

3.5. Approximate upper percentage points of Murphy's test statistic for two outliers

In this section we restrict our attention to the important case of $\nu = 0$, and illustrate that approximate percentage point of Murphy's statistic can be obtained from the tabulated percentage points of studentized range statistic T_{N6}. As shown in Section 3.1, the statistics U, M and T_{N3} are equivalent for $\nu = 0$. The relationships between the exact percentage points of these statistics are

$$(3.5.1) \qquad M_\alpha(e) = [2(n-2)/n]^{1/2} U_\alpha(e),$$

$$(3.5.2) \qquad T_{N3,\alpha}(e) = [2(n-1)(n-2)/n]^{1/2} U_\alpha(e),$$

where $M_\alpha(e)$, $T_{N3,\alpha}(e)$ and $U_\alpha(e)$ are the exact percentage points of M, T_{N3} and U respectively. The percentage points $M_\alpha(e)$ have been tabulated by Hawkins (1978) by evaluating a complicated integral for selected values of $n \leq 30$. Simulated percentage points of T_{N3} have been tabulated by Barnett and Lewis (1978) for n = 5(1)10(2)20(10)50,100. As mentioned in Section 3.3, the nominal percentage point u_α of U is exact for $n \leq 10$ and $\alpha = 0.05$. For smaller values of α, it is exact for slightly larger values of n.

Next consider the statistics V and T_{N6}. The relationship between the exact percentage points of these two statistics is obtained from (3.1.4) and is given by

$$(3.5.3) \qquad T_{N6,\alpha}(e) = [2(n-1)]^{1/2} V_\alpha(e),$$

where $V_\alpha(e)$ and $T_{N6,\alpha}(e)$ are the exact percentage points of V and T_{N6} respectively. The percentage points $T_{N6,\alpha}(e)$ were originally tabulated by David, Hartley and Pearson (1954) for selected values of n upto 1000. Their table has been extended by Pearson and Stephens (1964), and an abridged table is reproduced in Pearson and Hartley (1970).

The nominal percentage points u_α and v_α actually provide an upper limit to the exact values $U_\alpha(e)$ and $V_\alpha(e)$ respectively. From the calculations performed in Section 3.3 and 3.4, we see that u_α and v_α are quite close to the true percentage points for small values of n. Further, from equations (2.3.1) and (2.3.2) it follows that

(3.5.4) $$u_\alpha = v_{2\alpha}.$$

Since both u_α and $v_{2\alpha}$ are upper limits and are close to the true values, hence we expect that the exact percentage points of U and V are approximately related by an equation similar to (3.5.4). Thus, we are led to an approximation

(3.5.5) $$U_\alpha(a) \simeq V_{2\alpha}(e),$$

where $U_\alpha(a)$ stands for an approximate value of $U_\alpha(e)$. Now using equations (3.5.1), (3.5.2), (3.5.3) and (3.5.5) we immediately get

(3.5.6) $$U_\alpha(a) \simeq [1/\{2(n-1)\}^{1/2}] \cdot T_{N6,2\alpha}(e),$$

(3.5.7) $$T_{N3,\alpha}(a) \simeq [(n-2)/n]^{1/2} \cdot T_{N6,2\alpha}(e),$$

$$(3.5.8) \quad M_{\alpha}(a) \cong [(n-2)/\{n(n-1)\}]^{1/2} \cdot T_{N6,2\alpha}(e),$$

where $T_{N3,\alpha}(a)$ and $M_{\alpha}(a)$ stand for approximate value of $T_{N3,\alpha}(e)$ and $M_{\alpha}(e)$ respectively. To study the behaviour of this approximation, we compare the tabulated $M_{\alpha}(e)$ values with the approximate values obtained from equation (3.5.8). Unfortunately, due to limited tabulation of $T_{N6,2\alpha}(e)$ and $M_{\alpha}(e)$, we can do it only for $\alpha = 0.05$ and selected values of n. To extend our comparison, we also compare $T_{N3,\alpha}(a)$ with the simulated values $T_{N3,\alpha}(s)$ of T_{N3} as given by Barnett and Lewis (1978). These values are tabulated in Table 3.5.1, for n = 10(1) 20(5) 50(10)100, 200, 500 and 1000. In this table entries in columns corresponding to $T_{N6,2\alpha}(e)$, $T_{N3,\alpha}(s)$ and $M_{\alpha}(e)$ are taken from Pearson and Stephens (1964), Barnett and Lewis (1978) and Hawkins (1978) respectively. As can be seen that the approximation is remarkably good for almost all values of n for which the exact or simulated values are available. Some discrepancy could be due to the fact that $T_{N6,2\alpha}(e)$ is available only upto three significant digits. Consequently, the last digit in $T_{N3,\alpha}(a)$ and $M_{\alpha}(a)$ may be in error. Further, due to sampling variation, the simulated value $T_{N3,\alpha}(s)$ may not be equal to the exact value.

In view of such a close agreement, we recommend the use of approximate values for statistics M and T_{N3} as given in Table 3.5.1, whenever the exact values are not available, especially for large values of n.

In this study, we have proposed the statistic U, which of course is related to M and T_{N3}. We therefore provide a table of approximate percentage points $U_\alpha(a)$ obtained from equation (3.5.6). Again due to limited $T_{N6,2\alpha}(e)$ values available in Pearson and Stephens (1964), we tabulate $U_\alpha(a)$ for n = 10(1) 20(10) 100, 200, 500, 1000 and α = 0.005, 0.025 and 0.05 only. From the observations made above, $U_\alpha(a)$ is expected to be quite close to the true value $U_\alpha(e)$. Thus, for Murphy's test we need not use the nominal percentage points u_α, but can use approximate percentage points $U_\alpha(a)$. It should however be remembered that while u_α restricts the probability of type I error to α, we cannot make any such claim about $U_\alpha(a)$.

TABLE 3.3.2. Nominal upper percentile points of U and Murphy's statistic M.

ν	0				5			
n | α	0.01		0.05		0.01		0.05	
	u_α	M_α	u_α	M_α	u_α	M_α	u_α	M_α
5	0.9859	1.080	0.9587	1.050	0.8467	0.928	0.7646	0.838
6	0.9700	1.120	0.9326	1.077	0.8364	0.966	0.7598	0.877
7	0.9518	1.138	0.9074	1.085	0.8251	0.986	0.7523	0.899
8	0.9330	1.143	0.8840	1.083	0.8137	0.997	0.7436	0.911
9	0.9144	1.140	0.8624	1.076	0.8023	1.001	0.7344	0.916
10	0.8964	1.134	0.8424	1.066	0.7911	1.001	0.7251	0.917
11	0.8792	1.125	0.8239	1.054	0.7802	0.998	0.7158	0.916
12	0.8628	1.114	0.8069	1.042	0.7696	0.994	0.7067	0.912
13	0.8472	1.102	0.7910	1.029	0.7595	0.988	0.6978	0.908
14	0.8325	1.090	0.7762	1.016	0.7496	0.982	0.6892	0.902
15	0.8186	1.078	0.7623	1.004	0.7402	0.974	0.6808	0.896
20	0.7588	1.018	0.7044	0.945	0.6978	0.936	0.6430	0.863
25	0.7115	0.965	0.6598	0.895	0.6624	0.899	0.6112	0.829
30	0.6730	0.920	0.6239	0.852	0.6324	0.864	0.5841	0.798
35	0.6409	0.880	0.5941	0.816	0.6064	0.833	0.5607	0.770
40	0.6135	0.846	0.5688	0.784	0.5837	0.805	0.5402	0.745
45	0.5897	0.815	0.5470	0.756	0.5637	0.779	0.5220	0.722
50	0.5688	0.788	0.5279	0.731	0.5458	0.756	0.5058	0.701
55	0.5503	0.764	0.5109	0.709	0.5297	0.735	0.4913	0.682
60	0.5337	0.742	0.4957	0.689	0.5151	0.716	0.4780	0.665

TABLE 3.3.3. Nominal and tabulated critical values of T_{N3}.

n | α	0.05		0.01	
	Nominal	Tabulated	Nominal	Tabulated
10	3.192	3.18	3.397	3.40
20	4.113	4.11	4.431	4.41
50	5.120	5.06	5.517	5.51
100	5.743	5.62	6.164	6.06

TABLE 3.3.4. Nominal upper percentile points of V and T_{N6} for $\nu = 0$.

n | α	0.01		0.05	
	V_α	T_{N6}	V_α	T_{N6}
3	1.0000	2.000	0.9997	1.999
4	0.9983	2.445	0.9917	2.429
5	0.9911	2.803	0.9740	2.755
6	0.9788	3.095	0.9625	3.012
7	0.9636	3.338	0.9302	3.222
8	0.9470	3.543	0.9085	3.399
9	0.9301	3.720	0.8879	3.562
10	0.9133	3.875	0.8686	3.685
11	0.8970	4.012	0.8504	3.803
12	0.8813	4.134	0.8335	3.909
13	0.8663	4.244	0.8176	4.005
14	0.8519	4.344	0.8026	4.093
15	0.8382	4.435	0.7886	4.173
16	0.8251	4.519	0.7754	4.247
17	0.8127	4.597	0.7630	4.316
18	0.8008	4.669	0.7512	4.380
19	0.7894	4.737	0.7400	4.440
20	0.7786	4.800	0.7294	4.496
30	0.6917	5.268	0.6461	4.921
40	0.6308	5.571	0.5889	5.201
50	0.5849	5.790	0.5462	5.407
60	0.5487	5.960	0.5126	5.568
70	0.5191	6.098	0.4852	5.700
80	0.4943	6.213	0.4623	5.811
90	0.4731	6.311	0.4427	5.906
100	0.4546	6.397	0.4257	5.990

TABLE 3.4.1. Lower limits for type I error probability by using u_α for the statistic U.

ν	0		5	
n|α	0.01	0.05	0.01	0.05
5	0.01000	0.05000	0.01000	0.05000
6	0.01000	0.05000	0.01000	0.04999
7	0.01000	0.05000	0.01000	0.04994
8	0.01000	0.05000	0.01000	0.04983
9	0.01000	0.05000	0.01000	0.04965
10	0.01000	0.05000	0.01000	0.04942
11	0.01000	0.05000	0.01000	0.04915
12	0.01000	0.04999	0.00999	0.04885
13	0.01000	0.04993	0.00999	0.04852
14	0.01000	0.04980	0.00998	0.04817
15	0.01000	0.04961	0.00996	0.04781
20	0.00997	0.04805	0.00986	0.04585
25	0.00987	0.04602	0.00971	0.04383
30	0.00972	0.04392	0.00953	0.04184
35	0.00953	0.04183	0.00936	0.03997
40	0.00936	0.03991	0.00914	0.03800
45	0.00913	0.03794	0.00898	0.03642
50	0.00888	0.03619	0.00856	0.03472

TABLE 3.4.2. Lower limits for type I error probability by using v_α for the statistic V.

n \ α	0.01	0.05
10	0.01000	0.05000
12	0.01000	0.04996
14	0.01000	0.04970
16	0.01000	0.04924
18	0.00999	0.04864
20	0.00996	0.04798
30	0.00973	0.04447
40	0.00933	0.04105
50	0.00909	0.03801

TABLE 3.5.1. Comparison of approximate and exact percentage points of the Murphy's statistic for two outliers for $\alpha = 0.05$.

n	$T_{N6,2\alpha}(e)$	$T_{N3,\alpha}(a)$	$T_{N3,\alpha}(s)$	$M_\alpha(a)$	$M_\alpha(e)$
10	3.57	3.193	3.18	1.064	1.066
11	3.68	3.329		1.053	1.055
12	3.78	3.451	3.44	1.040	1.043
13	3.87	3.560		1.028	1.030
14	3.95	3.657	3.66	1.014	1.018
15	4.02	3.742		1.000	1.005
16	4.09	3.826	3.83	0.988	
17	4.15	3.898		0.975	
18	4.21	3.969	3.96	0.963	
19	4.27	4.039		0.952	
20	4.32	4.098	4.11	0.940	0.944
25	4.53	4.345		0.887	0.893
30	4.70	4.541	4.56	0.843	0.848
35	4.84	4.700		0.806	
40	4.96	4.834	4.84	0.774	
45	5.06	4.946		0.746	
50	5.14	5.036	5.06	0.719	
60	5.29	5.201		0.677	
70	5.41	5.332		0.642	
80	5.51	5.441		0.612	
90	5.60	5.537		0.587	
100	5.68	5.623	5.62	0.565	
200	6.15	6.119		0.434	
500	6.72	6.706		0.300	
1000	7.11	7.103		0.225	

TABLE 3.5.2. Approximate upper percentile points $U_\alpha(a)$ of the statistic U.

n | α	0.05		0.025		0.005	
	$T_{N6,2\alpha}(e)$	$U_\alpha(a)$	$T_{N6,2\alpha}(e)$	$U_\alpha(a)$	$T_{N6,2\alpha}(e)$	$U_\alpha(a)$
10	3.57	0.841	3.685	0.869	3.875	0.913
11	3.68	0.823	3.80	0.850	4.012	0.897
12	3.78	0.806	3.91	0.834	4.134	0.881
13	3.87	0.790	4.00	0.817	4.244	0.866
14	3.95	0.775	4.09	0.802	4.34	0.851
15	4.02	0.760	4.17	0.788	4.44	0.839
16	4.09	0.747	4.24	0.774	4.52	0.825
17	4.15	0.734	4.31	0.762	4.60	0.813
18	4.21	0.722	4.37	0.749	4.67	0.801
19	4.27	0.712	4.43	0.738	4.74	0.790
20	4.32	0.701	4.49	0.728	4.80	0.779
30	4.70	0.617	4.89	0.642	5.26	0.691
40	4.96	0.562	5.16	0.584	5.56	0.630
50	5.14	0.519	5.35	0.540	5.77	0.583
60	5.29	0.487	5.51	0.507	5.94	0.547
70	5.41	0.461	5.63	0.479	6.07	0.517
80	5.51	0.438	5.73	0.456	6.18	0.492
90	5.60	0.420	5.82	0.436	6.27	0.470
100	5.68	0.404	5.90	0.419	6.36	0.452
200	6.15	0.308	6.39	0.320	6.84	0.343
500	6.72	0.213	6.94	0.220	7.42	0.235
1000	7.11	0.159	7.33	0.164	7.80	0.175

CHAPTER IV

APPLICATION TO A TWO-WAY LAYOUT

4.1. Introduction

Two outliers in a random sample from $N(\mu,\sigma^2)$, discussed in Chapter III have been studied in maximum detail. After this, the most widely studied case is the detection of outliers in a two-way layout having a single observation in each cell. For example, Gentleman and Wilk (1975a), John and Draper (1978), Galpin and Hawkins (1981), Bradu and Hawkins (1982) etc. have discussed detection of one or more outliers in two-way tables. In this chapter, we will apply the general theory discussed in Chapter II to this case. For convenience we restrict our attention to $\nu = 0$ and use double subscripts in this chapter.

In Section 4.2, we derive the statistics for detecting two outliers. We then analyse the shape parameter existing between different $u_{i_1 j_1, i_2 j_2}$'s or $v_{i_1 j_1, i_2 j_2}$'s in Section 4.3, and discuss how these quantities can be used for finding bounds for exact percentile points. Finally in Section 4.4 we obtain the percentile points by Monte Carlo method and compare them with the nominal upper percentile points obtained in Section 2.3.

4.2. Test statistics

The model for a two-way layout with r rows and c columns and single observation in each cell is

(4.2.1) $E(Y_{ij}) = \mu + \alpha_i + \gamma_j$; $i = 1,2,\ldots,r$; $j = 1,2,\ldots,c$;

where μ is the general mean, α_i is the ith row effect and γ_j is the jth column effect. The total number of observations are thus $r.c = n$.

Rewriting (4.2.1) in the usual linear model form, we have

$$E(\underset{\sim}{Y}) = \underset{\sim}{X}\,\underset{\sim}{\beta},$$

where

$$\underset{n\times 1}{\underset{\sim}{Y}} = \begin{bmatrix} Y_{11} \\ Y_{12} \\ \vdots \\ Y_{1c} \\ \vdots \\ Y_{r1} \\ \vdots \\ Y_{rc} \end{bmatrix}, \qquad \underset{(r+c+1)\times 1}{\underset{\sim}{\beta}} = \begin{bmatrix} \mu \\ \alpha_1 \\ \vdots \\ \alpha_r \\ \gamma_1 \\ \vdots \\ \gamma_c \end{bmatrix}$$

and $\underset{\sim}{X}$ is the design matrix having the rank $k = r+c-1$.

The residuals for this model are

$$e_{ij} = Y_{ij} - Y_{i.} - Y_{.j} + Y_{..}, \quad i = 1,2,\ldots,r; \quad j = 1,2,\ldots,c;$$

where

Y_{ij} is the observation in (i,j)th cell,

$Y_{i.} = \sum_j Y_{ij}/c$,

$Y_{.j} = \sum_i Y_{ij}/r$, and

$Y_{..} = \sum_i \sum_j Y_{ij}/rc$.

The error sum of squares is

$$S^2 = \Sigma\Sigma\, e_{ij}^2 ,$$

which is also equal to S_p^2 as we are only considering the case $\nu = 0$. The degrees of freedom for S^2 are

$p = n-k = rc-(r+c-1) = (r-1)\,(c-1)$.

The correlation coefficient between any two residuals e_{ij} and $e_{i_1 j_1}$ depends upon their position in the table. It is given by

$$(4.2.2) \qquad \rho(e_{ij}, e_{i_1 j_1}) = R(ij, i_1 j_1) = \begin{cases} r_1 & \text{if } i \neq i_1 \text{ and } j \neq j_1 \\ r_2 & \text{if } i = i_1 \text{ and } j \neq j_1 \\ r_3 & \text{if } i \neq i_1 \text{ and } j = j_1 , \end{cases}$$

where $r_1 = 1/[(r-1)(c-1)]$, $r_2 = -1/(c-1)$ and $r_3 = -1/(r-1)$, and $R(ij, i_1 j_1)$ is used for notational convenience. Further the variance of each residual is $\lambda = (r-1)(c-1)/rc$.

Expressions for U and V become complicated due to double suffix notation. However, U is the maximum of $\binom{n}{2} = \binom{rc}{2}$ random variables and is given by

$$U = \text{Max}\,[\{u_{ij, i_1 j_1} : 1 \leq i = i_1 \leq r,\ 1 \leq j < j_1 \leq c\},$$

$$\{u_{ij, i_1 j_1} : 1 \leq i < i_1 \leq r,\ 1 \leq j \leq c,\ 1 \leq j_1 \leq c\}] .$$

Similar expression for V also holds.

For this case, the random variables needed for defining statistics U and V are

$$(4.2.3) \qquad u_{ij, i_1 j_1} = \frac{1}{S\,\lambda^{1/2}} \left[\frac{e_{ij} + e_{i_1 j_1}}{\{2(1+\rho(e_{ij}, e_{i_1 j_1}))\}^{1/2}} \right]$$

and

$$(4.2.4)\quad v_{ij,i_1j_1} = \frac{1}{s\,\lambda^{1/2}}\left[\frac{e_{ij} - e_{i_1j_1}}{\{2(1-\rho(e_{ij},e_{i_1j_1}))\}^{1/2}}\right],$$

$$i,i_1 = 1,2,\ldots,r \text{ and } j,j_1 = 1,2,\ldots,c.$$

In actual practice one does not have to evaluate all these quantities. Usually, calculations for few residuals having extreme values are sufficient. This has been illustrated in Example 2.1.1 of Section 2.1. It is observed that calculations for four or five residuals give U. Similarly, calculations for largest three and smallest three residuals are sufficient for V in most cases. Only in some extreme cases, one may have to do additional calculations.

4.3. Calculation of shape parameters

For finding the joint distribution of $u_{i_1j_1,i_2j_2}$ and $u_{i_3j_3,i_4j_4}$ or of $v_{i_1j_1,i_2j_2}$ and $v_{i_3j_3,i_4j_4}$, it is sufficient to consider the four residuals which are involved in defining these quantities. Gentleman (1980) has shown that there are 60 such combinations having distinct variance covariance matrix in a two-way table with r and c greater than or equal to 4. For other values of r and c, the number of these matrices is less than 60. These matrices can be identified with appropriate binary numbers or with their corresponding decimal numbers. The decimal numbers for these 60 matrices are given by Gentleman. In Table 4.3.1, we reproduce these decimal numbers as follows :

Suppose $\underset{2}{R}$ denotes the correlation matrix of the residuals $e_{i_1 j_1}$, $e_{i_2 j_2}$, $e_{i_3 j_3}$ and $e_{i_4 j_4}$, $i_1, i_2, i_3, i_4 = 1,2,\ldots,r$ and $j_1, j_2, j_3, j_4 = 1,2,\ldots,c$. Gentleman (1980) has shown that a decimal number can be used to represent a variance-covariance matrix. Same method can be extended to the correlation matrix $\underset{2}{R}$, since variances of all residuals are equal. We describe the procedure of Gentleman briefly. A given $\underset{2}{R}$ can be uniquely represented as two 4x4 binary matrices which will be denoted by $\underset{2}{H}$ and $\underset{2}{J}$. Denoting the 4 cell subscripts as $(i_1, j_1),\ldots,(i_4, j_4)$, the (a,b) element of $\underset{2}{H}$ and $\underset{2}{J}$ are defined as follows :

$$H(a,b) = \begin{cases} 1 & \text{if } i_a = i_b, \\ 0 & \text{otherwise,} \end{cases}$$

$$J(a,b) = \begin{cases} 1 & \text{if } j_a = j_b, \\ 0 & \text{otherwise,} \end{cases}$$

where $a = 1,2,\ldots,4$; $b = 1,2,\ldots,4$. As an example, suppose we want to express the correlation matrix of $(e_{13}, e_{21}, e_{22}, e_{23})$ in decimal form. The matrix $\underset{2}{R}$ can be written down immediately by using equation (4.2.2). Thus

$$\underset{2}{R} = \begin{bmatrix} 1 & r_1 & r_1 & r_3 \\ r_1 & 1 & r_2 & r_2 \\ r_1 & r_2 & 1 & r_2 \\ r_3 & r_2 & r_2 & 1 \end{bmatrix}.$$

Further, the matrices H_2 and J_2 are given by

$$H_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{bmatrix}, \quad J_2 = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}.$$

Since H_2 and J_2 are symmetric and all diagonal elements of both the matrices are necessarily 1, a given R_2 can be uniquely represented by a binary number consisting of 12 binary digits in the lower triangles of H_2 and J_2 (omitting diagonals), ordered, say, by rows, with elements of H_2 preceeding those of J_2. Thus the binary number corresponding to the example considered above is 001011000100. For the sake of brevity, this binary number is then converted to the corresponding decimal number. For the example cited above the decimal number is 708.

The method for writing R_2 from a given decimal number is just the reverse. We first calculate the corresponding 12 digit binary number, and obtain the matrices H_2 and J_2. For the (a,b) element of R_2 then we have

$$R(a,b) = \begin{cases} 1 & \text{if } H(a,b) = 1, \; J(a,b) = 1, \\ r_1 & \text{if } H(a,b) = 0, \; J(a,b) = 0, \\ r_2 & \text{if } H(a,b) = 1, \; J(a,b) = 0, \\ r_3 & \text{if } H(a,b) = 0, \; J(a,b) = 1 . \end{cases}$$

As an illustration, for the decimal number 530 given in Table 4.3.1, we have the binary number equal to 001000010010, and the matrices H_2 and J_2 as

$$H_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad J_2 = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}.$$

Finally,

$$R_2 = \begin{bmatrix} 1 & r_1 & r_3 & r_1 \\ r_1 & 1 & r_2 & r_3 \\ r_3 & r_2 & 1 & r_1 \\ r_1 & r_3 & r_1 & 1 \end{bmatrix}.$$

This, for example, is the correlation matrix of $(e_{12}, e_{21}, e_{22}, e_{31})$.

The entire work can be accomplished conveniently on a computer. Gentleman (1980) has not counted the number of matrices of each type. Although not directly needed for our study, yet it has been done for 3x3 and 4x5 layout, by generating all possible $\binom{rc}{4}$ combinations of 4 residuals for each case and obtaining the corresponding correlation matrix. For each matrix, corresponding decimal number was obtained by the process described above. The decimal numbers were then counted to give the desired frequencies. These are also reported in Table 4.3.1. Note that for a 3x3 layout, there are only 38 decimal numbers. In other words, if we consider all possible $\binom{9}{4} = 126$ combinations of 4 residuals, and obtain their correlation matrices, then there are only 38 such matrices. In addition $r = c = 3$ implies that $r_2 = -1/(c-1) = -0.5$ and $r_3 = -1/(r-1) = -0.5$. This in turn reduces the number of distinct correlation matrices to only 32. In this case, correlation matrices corresponding

to 6 pairs of decimal numbers (72,513), (96,2049), (120,3585), (544,2056), (545,2120) and (736,2059) are identical. Thus essentially we have only 32 distinct correlation matrices for this case.

4.3.1. Shape parameters of the bivariate distribution of $u_{i_1j_1,i_2j_2}$ and $u_{i_3j_3,i_4j_4}$

The shape parameter plays an important role for the determination of bivariate probabilities. In the present case, for the distribution of $u_{i_1j_1,i_2j_2}$ and $u_{i_3j_3,i_4j_4}$, it is given by

$$(4.3.1)\quad \rho_1(u_{i_1j_1,i_2j_2}, u_{i_3j_3,i_4j_4}) = \frac{R(i_1j_1,i_3j_3)+R(i_1j_1,i_4j_4)+R(i_2j_2,i_3j_3)+R(i_2j_2,i_4j_4)}{2[\{1+R(i_1j_1,i_2j_2)\}\{1+R(i_3j_3,i_4j_4)\}]^{1/2}},$$

where $i_1,i_2,i_3,i_4 = 1,2,\ldots,r$; $j_1,j_2,j_3,j_4 = 1,2,\ldots,c$, and $R(i_1j_1,i_2j_2)$ etc. are defined at equation (4.2.2).

Now for defining $u_{i_1j_1,i_2j_2}$ and $u_{i_3j_3,i_4j_4}$, we need 4 residuals. The correlation matrix $\underset{\sim}{R}$ of these contain $\binom{4}{2} = 6$ distinct correlations. Out of these six, two occur in the denominator of the expression for the shape parameter given in equation (4.3.1). Thus any correlation matrix $\underset{\sim}{R}$ of 4 residuals can give rise to at most $\binom{6}{2} = 15$ different shape parameters. Hence, there are at most 900 values of the shape parameters among the 60 different matrices. However, out of these 900 values only 40 are distinct. Out of these 40,

some may not exist or may coincide for special values of r and c. We shall denote them by ρ_i $(i = 1,2,\ldots,40)$. Also, since there are $N = \binom{n}{2}$ distinct $u_{i_1 j_1, i_2 j_2}$'s hence the number of bivariate distributions and shape parameters involved are $\binom{N}{2}$. Consequently, for 3x3, 4x5 and 5x6 layouts the total number of shape parameters are 630, 17955 and 94395 respectively. Different types of combinations of $u_{i_1 j_1, i_2 j_2}$ and $u_{i_3 j_3, i_4 j_4}$, corresponding expressions for shape parameters ρ in terms of r_1, r_2 and r_3, their frequencies for 3x3, 4x5 and 5x6 layouts and numerical values of ρ for these cases are tabulated in Table 4.3.2.

When r = c = 3, 13 shape parameters ρ_i, for i = 1,2,3, 4,5,17,18,23,24,32,34,39 and 40 do not exist. Out of remaining 27, only 11 are distinct. The ρ_i's which merge together in sets are $\{\rho_7, \rho_8, \rho_{25}, \rho_{26}, \rho_{29}, \rho_{30}, \rho_{35}, \rho_{36}\}$; $\{\rho_{10}, \rho_{11}, \rho_{21}, \rho_{22}\}$; $\{\rho_{12}, \rho_{19}, \rho_{20}\}$; $\{\rho_{13}, \rho_{14}\}$; $\{\rho_{15}, \rho_{16}\}$; $\{\rho_{31}, \rho_{33}\}$ and $\{\rho_{37}, \rho_{38}\}$.

For r = 4, c = 5, 37 of these ρ_i's are distinct. The ρ values having subscripts numbers 17, 19 and 21 in Table 4.3.2 merge with that of ρ_7 whose value is zero. For r = 5, c = 6, all 40 values of ρ_i's are different.

For r = c, we have $r_2 = r_3$ and several ρ_i's merge with other values. In this case there are only 19 distinct ρ_i values. The ρ_i values for the following sets of suffixes as given in Table 4.3.2 are identical :

{2,3,6}, {4,5}, {7,8}, {10,11}, {13,14}, {15,16}, {17,18}, {19,20}, {21,22}, {23,24,28}, {25,26,35,36}, {29,30}, {31,32,33,34}, {37,38}, {39,40}.

These values along with ρ_i for i = 1,9,12 and 27 constitute the 19 distinct values.

4.3.2. Shape parameters of the bivariate distribution of $v_{i_1j_1,i_2j_2}$ and $v_{i_3j_3,i_4j_4}$.

The shape parameter ρ' of the distribution of any two $v_{i_1j_1,i_2j_2}$ and $v_{i_3j_3,i_4j_4}$ is given by

$$(4.3.2)\quad \rho_1'(v_{i_1j_1,i_2j_2}, v_{i_3j_3,i_4j_4}) = \frac{R(i_1j_1,i_3j_3)-R(i_1j_1,i_4j_4)-R(i_2j_2,i_3j_3)+R(i_2j_2,i_4j_4)}{2[\{1-R(i_1j_1,i_2j_2)\}\{1-R(i_3j_3,i_4j_4)\}]^{1/2}},$$

where $i_1,i_2,i_3,i_4 = 1,2,\ldots,r$; $j_1,j_2,j_3,j_4 = 1,2,\ldots,c$ and $R(i_1j_1,i_2j_2)$ etc. are defined at equation (4.2.2).

For a two-way table, the total number of distinct ρ's is now 43. We label them as ρ_i' (i = 1,2,...,43). As said before, total number of bivariate distributions and shape parameters are $\binom{N}{2}$, where $N = \binom{n}{2}$.

Different types of combinations of $v_{i_1j_1,i_2j_2}$ and $v_{i_3j_3,i_4j_4}$, corresponding expressions for shape parameters ρ' in terms of r_1, r_2 and r_3, their frequencies of occurrence for 3x3, 4x5 and 5x6 layouts and numerical values of ρ' are given in Table 4.3.3. An asterisk sign in the frequency column

indicates that the same ρ'_i value has occured in some preceding row (shown in brackets) for some other combination and the frequency of it is merged with the earlier one. Here the frequencies of different combinations are difficult to count separately as was done for 'u' case, because in this case the order of each residual which constitutes $v_{i_1 j_1, i_2 j_2}$ and $v_{i_3 j_3, i_4 j_4}$ matters in evaluation of ρ'_i s. Note that for application purposes, we only need numerical values of ρ'_i s along with their frequencies, and an abridged table, similar to Table 3.2.4 of Chapter III is sufficient.

When r = c = 3, only 13 of the ρ'_i s are distinct. The rest of them merge with one or the other. For the case of r = 4, c = 5 only 39 out of 43 are distinct. Numerical values of $\rho'_3, \rho'_7, \rho'_{28}$ and ρ'_{37} are equal to ρ'_{14} , ρ'_{11} , ρ'_{34} and ρ'_{40} respectively for this case. When r = c > 4, there are 22 distinct shape parameters. In this case also, many of the ρ'_i s merged together. The ρ'_i values for the following sets of suffixes as given in Table 4.3.3 are identical : {2,3}, {4,5}, {6,7}, {8,10}, {9,19,21,23}, {11,12,13}, {15,16}, {17,18}, {24,25,34,35}, {26,27}, {29,30,33), {31,32}, {38,39}, {40,41}, {42,43}.

These values along with ρ'_i for i=1,14,20,22,28,36 and 37 constitute the 22 distinct values.

4.4. Percentile points

In Section 2.3, we have obtained tables of nominal upper percentile points for several values of n and k. These tables can be used here for values of r and c satisfying $3 \leq r, c \leq 11$ and $r+c \leq 14$. For some other values also we can use these tables. For comparison purposes, we now obtain lower and upper bounds for exact percentage points $U_\alpha(e)$ and $V_\alpha(e)$ for some selected values of α for two-way classification with $(r,c) = (4,5)$ and $(5,6)$. Percentile points by Monte Carlo procedure for some additional values of r and c have also been obtained.

As discussed in Section 2.4, bounds for the exact percentage point $U_\alpha(e)$ of U can be obtained by using the Bonferroni inequalities. An upper limit for $U_\alpha(e)$ is u_α, given in equation (2.3.1), while a lower limit $u_{*\alpha}$ for $U_\alpha(e)$ is obtained from equation (2.3.3). Due to bivariate probabilities involved, a direct solution of equation (2.3.3) is not possible. However, a recurring procedure could be adopted for small values of r and c. For the sake of notational simplicity, relabel all $u_{11,12}, u_{11,13}, \ldots, u_{r(c-1),rc}$ as u_i $(i = 1,2,\ldots,N)$, where $N = \binom{rc}{2}$. Then,

$$U = \max_{1 \leq i \leq N} u_i$$

and

$$S_1(u) - S_2(u) \leq \Pr(U \geq u) \leq S_1(u), \tag{4.4.1}$$

where

$$(4.4.2) \qquad S_1(u) = \binom{N}{2} \Pr(u_1 > u)$$

and

$$(4.4.3) \qquad S_2(u) = \sum_{1 \le i < j \le N} \Pr(u_i > u, u_j > u).$$

For finding a lower limit, we essentially have to solve $S_1(u) - S_2(u) = \alpha$. For this, we start with the upper limit value u_α. Corresponding to this u_α value, we first determine those bivariate probabilities appearing in equation (4.4.3), which are not equal to zero. If ρ is the shape parameter between u_i and u_j, then from equation (2.4.11), we see that

$$M(u_\alpha, u_\alpha, \rho, p) = \Pr(u_i > u_\alpha, u_j > u_\alpha) = 0 \text{ if } \rho \le 2u_\alpha^2 - 1 = \rho^* \text{ (say)}.$$

As an example, for $\alpha = 0.05$ and for a 4x5 table we get $\rho^* = 0.3593$. Now referring to Table 4.3.2, we see that bivariate probabilities corresponding to ρ_i for i = 9, 13, 14, 15 and 16 are non-zero. Thus five bivariate probabilities have to be evaluated. However, the number of bivariate probabilities to be evaluated increases rapidly as r and c increase. For example, for $\alpha = 0.05$, and r = 5, c = 6 we have $\rho^* = 0.0382$ and the number of bivariate probabilities to be evaluated increases to 16.

After evaluating all the non-zero bivariate probabilities, $M(u_\alpha, u_\alpha, \rho_i, p)$ we multiply each of them with their respective frequency, which is again given in Table 4.3.2. The sum of these

would then give $S_2(u_\alpha)$ and we obtain a lower limit of the significance probability, that is $\alpha - S_2(u_\alpha)$. It is then checked, whether this lower limit deviates from α by more than 10^{-4} (say). If yes, then let $\alpha_1 = \alpha + S_2(u_\alpha)$ and determine the corresponding u_{α_1} value. Again the lower limit of the significance probability, that is $\alpha_1 - S_2(u_{\alpha_1})$ is calculated and compared with α. This process is repeated until the lower limit for $Pr(U > u)$ does not differ from α by more than 10^{-4}. The value of u which gives this final lower limit of the significance probability can be considered as a lower bound for the true αth percentile point of U. We denote this value as $u_{*\alpha}$. Note that $u_{*\alpha}$, subject to calculation approximations, satisfies

$$Pr(U > u_{*\alpha}) \geq S_1(u_{*\alpha}) - S_2(u_{*\alpha}) = \alpha.$$

Similarly, the bounds for $V_\alpha(e)$ are also calculated. Here again we denote $v_{11,12}, v_{11,13}, \ldots, v_{r(c-1),rc}$ by $v_1, v_2, \ldots, v_N$, where $N = \binom{rc}{2}$ and

$$V = \max_{1 \leq i \leq N} |v_i|.$$

Now, we have

$$S_1(v) - S_2(v) \leq Pr(V \geq v) \leq S_1(v),$$

where $S_1(v) = \binom{N}{2} Pr(|v_1| > v)$ and

$$S_2(v) = \sum_{1 \leq i < j \leq N} Pr(|v_i| > v, |v_j| > v).$$

Similar to the case of U, we again begin with the nominal percentile value v_α. Now from equation (2.4.13) we have

$$\Pr(|v_i| > v_\alpha, |v_j| > v_\alpha) = 0 \text{ if } |\rho'| \leq \rho'^* = 2v_\alpha^2 - 1,$$

where ρ' is the shape parameter between v_i and v_j. If ρ'^* is negative, then of course, we have to evaluate all bivariate probabilities. But if ρ'^* is positive, then we have to evaluate only those bivariate probability terms for which $|\rho'| > \rho^*$. Note that

$$\Pr(|v_i| > v_\alpha, |v_j| > v_\alpha) = 2\left[M(v_\alpha, v_\alpha, \rho', p) + M(v_\alpha, v_\alpha, -\rho', p)\right],$$

and consequently, the value of this quantity for ρ' and $-\rho'$ is equal. For $\alpha = 0.05$, $r = 4$ and $c = 5$, we have $\rho'^* = 0.4331$. From Table 4.3.3, we see that the bivariate probabilities which have to be evaluated correspond to ρ'_i for $i = 1, 2, 4, 5, 8, 20, 22, 29, 36, 38$ and 39. Out of these, $\rho'_1, \rho'_4, \rho'_5, \rho'_8$ and ρ'_{20} are equal to $-\rho'_{36}$, $-\rho'_{38}$, $-\rho'_{39}$, $-\rho'_{29}$ and $-\rho'_{22}$ respectively. Thus essentially 6 bivariate probability terms like $\Pr(|v_i| > v_\alpha, |v_j| > v_\alpha)$ have to be calculated. This involves an evaluation of $M(v_\alpha, v_\alpha, \rho, p)$ probabilities for 12 different (both positive and negative) values of ρ. These are evaluated and then multiplied with suitable frequencies obtained from Table 4.3.3 and added to get $S_2(v_\alpha)$. The same iterative procedure as for $u_{*\alpha}$ is then followed for obtaining the bound $v_{*\alpha}$.

The bounds for $U_\alpha(e)$ for $(r, c) = (4, 5)$ and $(5, 6)$; $\alpha = 0.01$, 0.05 and 0.10 are given in Table 4.4.1. Similarly the bounds of $V_\alpha(e)$ for $r = 4$, $c = 5$ and for the same values of α are tabulated in Table 4.4.2.

Some percentile points for cases (r,c) = (3,3), (4,5), (5,4), (5,6), (6,5), (6,10) and (10,6) are obtained by Monte Carlo method also. These are given in Table 4.4.3 and Table 4.4.5 for the statistics U and V respectively.

The method followed for obtaining simulated percentile points was as follows. First rc standard normal variates for a rxc table were generated. Then the residuals were obtained in usual manner and $u_{i_1 j_1, i_2 j_2}$ values were calculated. Due to cost considerations $u_{i_1 j_1, i_2 j_2}$ values corresponding to five largest residuals were calculated. Their maximum then gives the value of statistic U. The process was repeated 1000 times to get 1000 values of statistic U. These 1000 values of U were arranged in descending order of magnitude to obtain an estimate of upper 100α percent point.

The entire process was repeated 25 times, and the average of these 25 values was calculated for obtaining simulated upper 100α percent point of U. Due to cost consideration, the average of only 15 repetitions was taken for (r,c) = (6,10) and (10,6) cases. These values are given in Table 4.4.3 for U for α = 0.005, 0.01, 0.025, 0.05 and 0.10. Due to symmetry, the percentile points of test statistic with r and c interchanged should remain unchanged. Consequently, the mean of the percentage points obtained for (r,c) and (c,r) combinations is also tabulated. This gives the desired percentage point by Monte Carlo method.

Exactly similar procedure was followed to obtain percentag points of V except that now $v_{i_1 j_1, i_2 j_2}$ values were calculated for all combinations of 3 largest and 3 smallest residuals only. Simulated percentage points for U denoted by $u_\alpha(s)$ and for V denoted by $v_\alpha(s)$ are given in Tables 4.4.4 and 4.4.6 respectively. Corresponding nominal percentage points obtained in Section 2.3 are also provided in parenthesis. As can be seen from Tables 4.4.4 and 4.4.6, for both U and V, nominal percentage points provide a fairly good approximation for true percentage points even for moderately large values of (r,c) and α. Thus for $r = 6$, $c = 10$ and $\alpha = 0.10$, simulated percentage point for U is $u_\alpha(s) = 0.5281$ while nominal percentage point is $u_\alpha = 0.5385$. Similarly for $r = 6$, $c = 10$ and $\alpha = 0.20$ we have $v_\alpha(s) = 0.5257$, while the nominal percentage point is $v_\alpha = 0.5385$.

Due to sampling variations, some of the simulated values ar outside the corresponding lower and upper bounds given in Table 4.4.1 and Table 4.4.2, especially for small values of α, which indicates that a much larger number of repetitions have to be performed for satisfactory percentage points by Monte Carlo method. Even otherwise, simulated percentage points usually have accuracy of only 2 or 3 significant digits. We therefore recommend the use of nominal percentage points u_α and v_α for U and V respectively, which are quite easy to evaluate.

TABLE 4.3.1. Sixty distinct matrices and their frequency of occurrence for a 3x3 and a 4x5 table.

Decimal number of the matrices	Frequency 3x3 table	Frequency 4x5 table	Decimal number of the matrices	Frequency 3x3 table	Frequency 4x5 tabl
0		120	530	3	40
1		60	533	3	40
2		60	544	3	120
4		60	545	3	40
8		60	550	3	40
11		20	704		120
12		20	708	3	60
16		60	720	3	60
18		20	736	3	60
21		20	2048		240
32		60	2049	3	120
33		20	2050	3	120
38		20	2052	3	120
56		20	2056	3	120
63		5	2059	3	40
64		240	2060	3	40
66	3	120	2064	3	120
68	3	120	2066	3	40
72	3	120	2069	3	40
76	3	40	2112		180
80	3	120	2114	6	120
82	3	40	2116	3	60
96	3	120	2120	3	60
102	3	40	2128	6	120
120	3	40	2130	9	60
512		240	3584		120
513	3	120	3585	3	60
514	3	120	3586	3	60
516	3	120	3588	3	60
528	3	120	4032		20
			Total	126 = $\binom{9}{4}$	4845 = $\binom{20}{4}$

TABLE 4.3.2. Different combinations of $u_{i_1 j_1, i_2 j_2}$'s with shape parameters and frequency of occurrence for 3x3, 4x5 and 5x6 tables[1].

Subscripts of ρ_i's	Type of combinations	Formula of the ρ_i's	Frequencies and the value the ρ_i's		
			3x3 table	4x5 table	5x6 table
1	$u_{11,22}, u_{33,44}$	$\frac{2r_1}{(1+r_1)}$	0	360 .1538	5400 .0952
2	$u_{11,12}, u_{23,24}$	$\frac{2r_1}{(1+r_2)}$	0	180 .2222	900 .1250
3	$u_{11,21}, u_{32,42}$	$\frac{2r_1}{(1+r_3)}$	0	60 .2500	450 .1333
4	$u_{11,22}, u_{33,34}$	$\frac{2r_1}{[(1+r_1)(1+r_2)]^{1/2}}$	0	720 .1849	5400 .1091
5	$u_{11,22}, u_{33,43}$	$\frac{2r_1}{[(1+r_1)(1+r_3)]^{1/2}}$	0	360 .1961	3600 .1127
6	$u_{11,12}, u_{23,33}$	$\frac{2r_1}{[(1+r_2)(1+r_3)]^{1/2}}$	9 1.0000	360 .2357	1800 .1291
7	$u_{11,21}, u_{11,31}$	$\frac{(1+3r_3)}{2(1+r_3)}$	9 -.5000	60 .0000	180 .1667
8	$u_{11,12}, u_{11,13}$	$\frac{(1+3r_2)}{2(1+r_2)}$	9 -.5000	120 .1667	300 .2500
9	$u_{11,22}, u_{11,33}$	$\frac{(1+3r_1)}{2(1+r_1)}$	18 .7000	720 .5769	3600 .5476
10	$u_{11,12}, u_{11,22}$	$\frac{(1+r_1+r_2+r_3)}{2[(1+r_1)(1+r_2)]^{1/2}}$	36 .1581	240 .2774	600 .3273

[1] The values shown in the bottom row are the numerical values of shape parameters.

TABLE 4.3.2.Contd.

Subscripts of ρ_i's	Type of combinations	Formula of the ρ_i's	Frequencies and the value of the ρ_i's		
			3x3 table	4x5 table	5x6 table
11	$u_{11,21}, u_{11,22}$	$\frac{(1+r_1+r_2+r_3)}{2[(1+r_1)(1+r_3)]^{1/2}}$	36 .1581	240 .2942	600 .3381
12	$u_{11,12}, u_{11,21}$	$\frac{(1+r_1+r_2+r_3)}{2[(1+r_2)(1+r_3)]^{1/2}}$	36 .2500	240 .3536	600 .3873
13	$u_{11,22}, u_{11,32}$	$\frac{(1+2r_1+r_3)}{2(1+r_1)}$	18 .4000	240 .3846	900 .4048
14	$u_{11,22}, u_{11,23}$	$\frac{(1+2r_1+r_2)}{2(1+r_1)}$	18 .4000	360 .4231	1200 .4286
15	$u_{11,22}, u_{22,32}$	$\frac{(1+2r_1+r_3)}{2[(1+r_1)(1+r_3)]^{1/2}}$	36 .6325	480 .4903	1800 .4789
16	$u_{11,22}, u_{22,23}$	$\frac{(1+2r_1+r_2)}{2[(1+r_1)(1+r_2)]^{1/2}}$	36 .6325	720 .5085	2400 .4910
17	$u_{11,22}, u_{23,34}$	$\frac{(3r_1+r_2)}{2(1+r_1)}$	0	1440 .0000	10800 -.0238
18	$u_{11,22}, u_{32,43}$	$\frac{(3r_1+r_3)}{2(1+r_1)}$	0	720 -.0385	7200 -.0476
19	$u_{11,21}, u_{22,32}$	$\frac{(3r_1+r_2)}{2(1+r_3)}$	18 .2500	240 .0000	900 -.0333
20	$u_{11,12}, u_{22,23}$	$\frac{(3r_1+r_3)}{2(1+r_2)}$	18 .2500	360 -.0556	1200 -.0625
21	$u_{11,22}, u_{23,33}$	$\frac{(3r_1+r_2)}{2[(1+r_1)(1+r_3)]^{1/2}}$	36 .1581	1440 .0000	7200 -.0282

TABLE 4.3.2. Contd.

Subscripts of $'\rho_i's$	Type of combinations	Formula of the $'\rho_i's$	Frequencies and the value of the $'\rho_i's$		
			3x3 table	4x5 table	5x6 table
22	$u_{11,22}, u_{32,33}$	$\frac{(3r_1+r_3)}{2[(1+r_1)(1+r_2)]^{1/2}}$	36 .1581	1440 .0462	7200 -.0546
23	$u_{11,22}, u_{13,24}$	$\frac{(r_1+r_2)}{(1+r_1)}$	0	360 -.1538	1800 -.1429
24	$u_{11,22}, u_{32,41}$	$\frac{(r_1+r_2)}{(1+r_1)}$	0	120 -.2308	900 -.1905
25	$u_{11,21}, u_{12,22}$	$\frac{(r_1+r_2)}{(1+r_3)}$	9 -.5000	60 -.2500	150 -.2000
26	$u_{11,12}, u_{21,22}$	$\frac{(r_1+r_3)}{(1+r_2)}$	9 -.5000	60 -.3333	150 -.2500
27	$u_{11,22}, u_{12,21}$	$\frac{(r_2+r_3)}{(1+r_1)}$	9 -.8000	60 -.5385	150 -.4286
28	$u_{11,32}, u_{22,33}$	$\frac{(2r_1+r_2+r_3)}{2(1+r_1)}$	54 -.2000	2160 -.1923	10800 -.1667
29	$u_{11,22}, u_{12,23}$	$\frac{(r_1+2r_2+r_3)}{2(1+r_1)}$	18 -.5000	360 -.3462	1200 -.2857
30	$u_{11,32}, u_{12,21}$	$\frac{(r_1+2r_3+r_2)}{2(1+r_1)}$	18 -.5000	240 -.3846	900 -.3095
31	$u_{11,23}, u_{12,22}$	$\frac{(r_1+r_2)}{[(1+r_1)(1+r_3)]^{1/2}}$	18 -.3162	360 -.1961	1200 -.1690

TABLE 4.3.2. Contd.

Subscripts of $'\rho_i's$	Type of combinations	Formula of the $'\rho_i's$	Frequencies and the value of the $'\rho_i's$		
			3x3 table	4x5 table	5x6 table
32	$u_{11,22}, u_{23,24}$	$\frac{(r_1+r_2)}{[(1+r_1)(1+r_2)]^{1/2}}$	0	720 -.1849	3600 -.1637
33	$u_{11,12}, u_{21,32}$	$\frac{(r_1+r_3)}{[(1+r_1)(1+r_2)]^{1/2}}$	18 -.3162	240 -.2774	900 -.2182
34	$u_{11,22}, u_{32,42}$	$\frac{(r_1+r_3)}{[(1+r_1)(1+r_3)]^{1/2}}$	0	240 -.2942	1800 -.2254
35	$u_{11,13}, u_{12,22}$	$\frac{(r_1+r_2)}{[(1+r_2)(1+r_3)]^{1/2}}$	18 -.5000	360 -.2357	1200 -.1936
36	$u_{11,31}, u_{21,22}$	$\frac{(r_1+r_3)}{[(1+r_2)(1+r_3)]^{1/2}}$	18 -.5000	240 -.3536	900 -.2582
37	$u_{11,23}, u_{12,13}$	$\frac{(r_1+2r_2+r_3)}{2[(1+r_1)(1+r_2)]^{1/2}}$	36 -.7906	720 -.4160	2400 -.3273
38	$u_{11,32}, u_{21,31}$	$\frac{(r_1+2r_3+r_2)}{2[(1+r_1)(1+r_3)]^{1/2}}$	36 -.7906	480 -.4903	1800 -.3662
39	$u_{11,12}, u_{13,14}$	$\frac{2r_2}{(1+r_2)}$	0	60 -.6667	225 -.5000
40	$u_{11,21}, u_{31,41}$	$\frac{2r_3}{(1+r_3)}$	0	15 -1.0000	90 -.6667
		Total	630	17955	94395

TABLE 4.3.3. Different combinations of $v_{i_1 j_1, i_2 j_2}$'s with shape parameters and frequency of occurrence for 3x3, 4x5 and 5x6 tables[1]

Subscripts ρ'_i	Type of combinations	Formula of the ρ'_i	Frequencies and the value of ρ'_i 3x3 table	4x5 table	5x6 table
1	$v_{11,12}, v_{11,21}$	$\frac{(1-r_2-r_3+r_1)}{2[(1-r_2)(1-r_3)]^{1/2}}$	18 .7500	120 .6455	300 .6124
2	$v_{11,21}, v_{11,22}$	$\frac{(1-r_1-r_3+r_2)}{2[(1-r_1)(1-r_3)]^{1/2}}$	114 .3536	240 .4523	600 .4588
3	$v_{11,12}, v_{11,22}$	$\frac{(1-r_1-r_2+r_3)}{2[(1-r_1)(1-r_2)]^{1/2}}$	*(2) .3536	240 .3892	300 .4215
4	$v_{11,32}, v_{22,32}$	$\frac{1-r_3}{2[(1-r_1)(1-r_3)]^{1/2}}$	51 .7071	320 .6030	1200 .5735
5	$v_{11,23}, v_{22,23}$	$\frac{1-r_2}{2[(1-r_1)(1-r_2)]^{1/2}}$	*(4) .7071	360 .5839	1200 .5620
6	$v_{11,22}, v_{11,23}$	$\frac{(1-2r_1+r_2)}{2(1-r_1)}$	126 .0000	360 .3182	1200 .3684
7	$v_{11,22}, v_{11,32}$	$\frac{(1-2r_1+r_3)}{2(1-r_1)}$	*(6) .0000	320 .2727	600 .3421
8	$v_{11,22}, v_{32,41}$	$\frac{(r_1-r_3)}{(1-r_1)}$	0	60 .4545	450 .3158
9	$v_{11,22}, v_{12,21}$	$\frac{(r_2-r_3)}{(1-r_1)}$	*(6) .0000	60 .0909	150 .0526

[1] The values shown in the bottom row are the numerical values of shape parameters. The frequency of those ρ'_i values denoted with an asterisk sign (*) indicates that the same value has occured in some preceeding row (shown within brackets), for some other combination and the frequency of it is merged with earlier frequency.

TABLE 4.3.3. Contd.

Subscripts ρ'_i	Type of combinations	Formula of the ρ'_i	Frequencies and the value of ρ'_i		
			3x3 table	4x5 table	5x6 table
10	$v_{11,22}, v_{23,31}$	$\frac{(2r_1-r_2-r_3)}{2(1-r_1)}$	6 1.0000	240 .4091	1200 .2895
11	$v_{11,32}, v_{12,21}$	$\frac{(r_1-2r_3+r_2)}{2(1-r_1)}$	36 .5000	*(7) .2727	600 .1842
12	$v_{11,22}, v_{32,43}$	$\frac{(r_1-r_3)}{2(1-r_1)}$	0	360 .2273	3600 .1579
13	$v_{11,22}, v_{23,34}$	$\frac{(r_1-r_2)}{2(1-r_1)}$	0	480 .1818	3600 .1316
14	$v_{11,12}, v_{22,31}$	$\frac{(r_1-r_3)}{[(1-r_1)(1-r_2)]^{1/2}}$	*(4) .7071	*(3) .3892	450 .2810
15	$v_{11,32}, v_{12,13}$	$\frac{(r_1-r_3)}{2[(1-r_1)(1-r_2)]^{1/2}}$	*(2) .3536	1080 .1946	4800 .1405
16	$v_{11,23}, v_{21,31}$	$\frac{(r_1-r_2)}{2[(1-r_1)(1-r_3)]^{1/2}}$	*(2) .3536	640 .1508	3000 .1147
17	$v_{11,12}, v_{22,23}$	$\frac{(r_1-r_3)}{2(1-r_2)}$	12 .2500	120 .1667	400 .1250
18	$v_{11,21}, v_{22,32}$	$\frac{(r_1-r_2)}{2(1-r_3)}$	*(17) .2500	80 .1250	300 .1000
19	$v_{11,32}, v_{22,33}$	$\frac{(r_2-r_3)}{2(1-r_1)}$	*(6) .0000	960 .0455	4800 .0263
20	$v_{11,21}, v_{11,31}$	$\frac{1}{2}$	*(11) .5000	600 .5000	2720 .5000

TABLE 4.3.3. Contd.

Subscripts ρ'_i	Type of combinations	Formula of the ρ'_i	Frequencies and the value of ρ'_i		
			3x3 table	4x5 table	5x6 table
21	$v_{11,22}, v_{33,44}$	0	0	3675 .0000	25365 .0000
22	$v_{11,12}, v_{12,13}$	$-\frac{1}{2}$	54 -.5000	300 -.5000	1360 -.5000
23	$v_{11,23}, v_{22,33}$	$-\frac{(r_2-r_3)}{2(1-r_1)}$	*(6) .0000	480 -.0455	2400 -.0263
24	$v_{11,22}, v_{31,43}$	$-\frac{(r_1-r_3)}{2(1-r_1)}$	0	360 -.2273	3600 -.1579
25	$v_{11,22}, v_{13,34}$	$-\frac{(r_1-r_2)}{2(1-r_1)}$	0	960 -.1818	7200 -.1316
26	$v_{11,12}, v_{21,23}$	$-\frac{(r_1-r_3)}{2(1-r_2)}$	24 -.2500	240 -.1667	800 -.1250
27	$v_{11,21}, v_{13,33}$	$-\frac{(r_1-r_2)}{2(1-r_3)}$	*(25) -.2500	160 -.1250	600 -.1000
28	$v_{11,22}, v_{22,31}$	$-\frac{(1-2r_1+r_3)}{2(1-r_1)}$	*(6) .0000	160 -.2727	300 -.3421
29	$v_{11,22}, v_{31,42}$	$-\frac{(r_1-r_3)}{(1-r_1)}$	0	60 -.4545	450 -.3158
30	$v_{11,22}, v_{13,24}$	$-\frac{(r_1-r_2)}{(1-r_1)}$	0	360 -.3636	1800 -.2632
31	$v_{11,12}, v_{21,22}$	$-\frac{(r_1-r_3)}{(1-r_2)}$	*(22) -.5000	60 -.3333	150 -.2500
32	$v_{11,21}, v_{12,22}$	$-\frac{(r_1-r_2)}{(1-r_3)}$	*(22) -.5000	60 -.2500	150 -.2000

TABLE 4.3.3..Contd.

Subscripts ρ'_i	Type of combinations	Formula of the ρ'_i	Frequencies and the value of ρ'_i 3x3 table	4x5 table	5x6 table
33	$v_{11,22}, v_{13,32}$	$-\frac{(2r_1-r_2-r_3)}{2(1-r_1)}$	12 -1.000	480 -.4091	2400 -.2895
34	$v_{11,22}, v_{21,32}$	$\frac{(2r_3-r_1-r_2)}{2(1-r_1)}$	*(22) -.5000	*(28) -.2727	300 -.1842
35	$v_{11,22}, v_{12,23}$	$\frac{(2r_2-r_1-r_3)}{2(1-r_1)}$	*(22) -.5000	360 -.1364	1200 -.1053
36	$v_{11,12}, v_{12,22}$	$\frac{-(1-r_2-r_3+r_1)}{2[(1-r_2)(1-r_3)]^{1/2}}$	18 -.7500	120 -.6455	300 -.6124
37	$v_{11,12}, v_{12,21}$	$\frac{-(1-r_1-r_2+r_3)}{2[(1-r_1)(1-r_2)]^{1/2}}$	102 -.3536	240 -.3892	300 -.4215
38	$v_{11,22}, v_{22,32}$	$\frac{-(1-r_3)}{2[(1-r_1)(1-r_3)]^{1/2}}$	57 -.7071	160 -.6030	600 -.5735
39	$v_{11,22}, v_{22,23}$	$\frac{-(1-r_2)}{2[(1-r_1)(1-r_2)]^{1/2}}$	*(38) -.7071	360 -.5839	1200 -.5620
40	$v_{11,23}, v_{31,32}$	$\frac{-(r_1-r_3)}{[(1-r_1)(1-r_2)]^{1/2}}$	*(38) -.7071	*(37) -.3892	450 -.2810
41	$v_{11,23}, v_{12,22}$	$\frac{-(r_1-r_2)}{[(1-r_1)(1-r_3)]^{1/2}}$	*(38) -.7071	360 -.3015	1200 -.2294
42	$v_{11,23}, v_{12,13}$	$\frac{-(r_1-r_3)}{2[(1-r_1)(1-r_2)]^{1/2}}$	*(37) -.3536	1080 -.1946	4800 -.1405
43	$v_{11,32}, v_{21,31}$	$\frac{-(r_1-r_2)}{2[(1-r_1)(1-r_3)]^{1/2}}$	*(37) -.3536	1280 -.1508	6000 -.1147
		Total :	630	17955	94395

TABLE 4.4.1. Bounds for $U_\alpha(e)$ for $(r,c) = (4,5)$ and $(5,6)$.

α	$r = 4,\ c = 5$		$r = 5,\ c = 6$	
	Lower bound	Upper bound	Lower bound	Upper bound
0.01	0.8712	0.8712	0.7689	0.7692
0.05	0.8242	0.8244	0.7181	0.7205
0.10	0.7978	0.7989	0.6898	0.6959

TABLE 4.4.2. Bounds for $V_\alpha(e)$ for $r=4$ and $c = 5$.

α	Lower bound	Upper bound.
0.01	0.88710	0.88710
0.05	0.84631	0.84647
0.10	0.82362	0.82443

TABLE 4.4.3. $u_\alpha(s)$ values obtained by Monte Carlo method.

r	c	α				
		0.005	0.01	0.025	0.05	0.10
3	3	0.9952	0.9927	0.9869	0.9791	0.9668
4	5	0.8859	0.8684	0.8449	0.8232	0.7965
5	4	0.8878	0.8729	0.8462	0.8245	0.7989
5	6	0.7878	0.7691	0.7405	0.7176	0.6889
6	5	0.7864	0.7671	0.7396	0.7168	0.6904
6	10	0.6095	0.5931	0.5692	0.5492	0.5273
10	6	0.6102	0.5952	0.5704	0.5515	0.5290

TABLE 4.4.4. $u_\alpha(s)$ values obtained by taking the average of (r,c) and (c,r) values of Table 4.4.3[1].

r	c	α				
		0.005	0.01	0.025	0.05	0.10
3	3	0.9952 (0.9962)	0.9927 (0.9940)	0.9869 (0.9890)	0.9791 (0.9825)	0.9668 (0.9722)
4	5	0.8869 (0.8871)	0.8706 (0.8712)	0.8456 (0.8465)	0.8238 (0.8244)	0.7977 (0.7989)
5	6	0.7871 (0.7871)	0.7681 (0.7692)	0.7400 (0.7428)	0.7172 (0.7205)	0.6896 (0.6959)
6	10	0.6098 (0.6141)	0.5941 (0.5982)	0.5698 (0.5758)	0.5503 (0.5578)	0.5281 (0.5385)

[1]The values shown in parenthesis are the nominal percentile points obtained in Section 2.3.

TABLE 4.4.5. $v_\alpha(s)$ values obtained by Monte Carlo method.

r	c	α				
		0.01	0.02	0.05	0.10	0.20
3	3	0.9951	0.9924	0.9859	0.9767	0.9628
4	5	0.8888	0.8722	0.8465	0.8229	0.7954
5	4	0.8857	0.8691	0.8443	0.8219	0.7943
5	6	0.7880	0.7674	0.7409	0.7156	0.6872
6	5	0.7823	0.7656	0.7406	0.7170	0.6878
6	10	0.6116	0.5959	0.5708	0.5507	0.5266
10	6	0.6128	0.5953	0.5712	0.5496	0.5248

TABLE 4.4.6. $v_\alpha(s)$ values obtained by taking the average of (r,c) and (c,r) values from Table 4.4.5[1].

r	c	α				
		0.01	0.02	0.05	0.10	0.20
3	3	0.9951 (0.9962)	0.9924 (0.9940)	0.9859 (0.9890)	0.9767 (0.9825)	0.9628 (0.9722)
4	5	0.8872 (0.8871)	0.8707 (0.8712)	0.8454 (0.8465)	0.8224 (0.8244)	0.7948 (0.7989)
5	6	0.7851 (0.7871)	0.7665 (0.7692)	0.7408 (0.7428)	0.7163 (0.7205)	0.6875 (0.6959)
6	10	0.6122 (0.6141)	0.5956 (0.5982)	0.5710 (0.5758)	0.5501 (0.5578)	0.5257 (0.5385)

[1]The values shown in parenthesis are the nominal percentile points obtained in Section 2.3.

CHAPTER V

PERFORMANCE OF THE STATISTICS

5.1. Introduction

In this chapter we will study the performance of the test statistics proposed in Section 2.1 in the non-null situation when two outliers are present. Similar to Chapter IV, we again consider the case $\nu = 0$ only. The distribution theory results for $\nu > 0$ are exactly analogous with minor changes at few places. Consequently, we shall use S^2 in place of S_p^2. The statistic u_{ij} of equation (2.1.1) then reduces to

$$(5.1.1) \qquad u_{ij} = (e_i/\lambda_{ii}^{1/2} + e_j/\lambda_{jj}^{1/2})/[S\{2(1+\rho_{ij})\}^{1/2}] .$$

Similar expression holds for v_{ij}. Let $y_1, y_2, \ldots, y_n$ be n independent and normally distributed observations. We now assume that exactly two of these observations are outliers. The null hypothesis H_o is that there is no outlier, and under H_o the model is given by equation (1.2.1), viz. for $i = 1, 2, \ldots,$

$$E(Y_i) = \sum_{j=1}^{m} x_{ij} \beta_j = \mu_i, \text{ say,}$$

$$\text{Var } (Y_i) = \sigma^2.$$

To evaluate the performance of the statistic U we use a one-sided alternative hypothesis, and for V a two-sided alternati hypothesis is used. For one-sided hypothesis we assume that two observations, we do not know which ones, have a mean shifted to the right. The alternative hypothesis is then the union of

$\binom{n}{2}$ hypotheses H_{ij} $(1 \leq i < j \leq n)$, where under H_{ij}, we have

$$E(Y_s) = \begin{cases} \mu_s & \text{if } s \neq i,j, \\ \mu_i + \theta_i & \text{if } s = i, \\ \mu_j + \theta_j & \text{if } s = j, \end{cases}$$

where θ_i and θ_j are greater than zero. Other assumptions regarding the variance and distribution of $Y_1, Y_2, \ldots, Y_n$ remain unchanged. Consequently, under H_{ij}

$$\underset{\sim}{Y} \overset{d}{=} N(.,\sigma^2 \underset{\sim}{I}), \tag{5.1.2}$$

where

$$E(\underset{\sim}{Y}) = \underset{\sim}{X}\,\underset{\sim}{\beta} + \theta_i\,\underset{\sim}{\varepsilon}_i + \theta_j\,\underset{\sim}{\varepsilon}_j, \tag{5.1.3}$$

$\underset{\sim}{\varepsilon}_i$ $(i = 1,2,\ldots,n)$ is the ith column of $\underset{\sim}{I}_n$, the identity matrix of order n.

Similarly, for two-sided hypothesis, we assume that out of two observations, one observation has a mean shifted to left while the other one has a mean shifted to right. We discuss the distribution theory of u_{ij} and v_{ij} etc. under the alternative hypothesis in next two sections. Measures of performance of these test statistics U and V are studied in Section 5.4. Finally, we compare our procedure with that of sequential procedure suggested by Anscombe (1960), Moran and McMillan (1973), and John and Draper (1978), etc.

5.2. Non-null distribution of u_{ij}

For notational convenience we shall denote the random vector $\underset{\sim}{Y}$ by $\underset{\sim}{y}$ etc. From equation (1.2.2), the residual vector

is $\underset{\sim}{e} = \underset{\sim}{\Lambda}\, \underset{\sim}{y}$. Further, as stated in Section 1.2, the residual vector $\underset{\sim}{e}$ has a singular normal distribution $N(\underset{\sim}{0}, \underset{\sim}{\Lambda}\sigma^2)$ under H_0. We now obtain the distribution of $\underset{\sim}{e}$ and residual sum of squares S^2 under H_{ij}.

Clearly, under H_{ij}, $\underset{\sim}{e}$ is normally distributed with variance-covariance matrix $\underset{\sim}{\Lambda}\sigma^2$ and mean

$$\begin{aligned} E(\underset{\sim}{e}|H_{ij}) &= \underset{\sim}{\Lambda}\, E(\underset{\sim}{y} \mid H_{ij}) \\ &= \underset{\sim}{\Lambda}(\underset{\sim}{X}\,\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}_i\,\theta_i + \underset{\sim}{\varepsilon}_j\,\theta_j) \\ &= \underset{\sim}{\Lambda}(\underset{\sim}{\varepsilon}_i\,\theta_i + \underset{\sim}{\varepsilon}_j\,\theta_j), \end{aligned}$$

since $\underset{\sim}{\Lambda}\underset{\sim}{X} = \underset{\sim}{0}$.

The residual sum of squares S^2 has a non-central $\sigma^2\chi^2$ distribution with n-k degrees of freedom and non-centrality parameter λ^*, where

$$\begin{aligned} \sigma^2\lambda^* &= E(\underset{\sim}{y}'|H_{ij})\ \underset{\sim}{\Lambda}\ E(\underset{\sim}{y}|H_{ij}) \\ &= (\underset{\sim}{\beta}'\,\underset{\sim}{X}' + \underset{\sim}{\varepsilon}_i'\,\theta_i + \underset{\sim}{\varepsilon}_j'\,\theta_j)\ \underset{\sim}{\Lambda}\ (\underset{\sim}{X}\,\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}_i\,\theta_i + \underset{\sim}{\varepsilon}_j\,\theta_j) \\ &= \underset{\sim}{\varepsilon}_i'\underset{\sim}{\Lambda}\,\underset{\sim}{\varepsilon}_i\,\theta_i^2 + \underset{\sim}{\varepsilon}_j'\,\underset{\sim}{\Lambda}\,\underset{\sim}{\varepsilon}_j\,\theta_j^2 + \underset{\sim}{\varepsilon}_i'\underset{\sim}{\Lambda}\,\underset{\sim}{\varepsilon}_j\theta_i\theta_j + \underset{\sim}{\varepsilon}_j'\underset{\sim}{\Lambda}\,\underset{\sim}{\varepsilon}_i\,\theta_i\,\theta_j \\ (5.2.1) &= \lambda_{ii}\,\theta_i^2 + \lambda_{jj}\,\theta_j^2 + 2\,\lambda_{ij}\,\theta_i\,\theta_j. \end{aligned}$$

We denote this as $S^2 \overset{d}{=} \sigma^2\chi^2(n-k,\lambda^*)$.

For the sake of convenience we consider H_{12}, and derive the distribution of u_{ij} under H_{12}. For this, we need the following properties of $\underset{\sim}{\Lambda}$.

Property (i) : $\underset{\sim}{\Lambda}$ is a symmetric and idempotent matrix of rank (n-k).

Property (ii) : Let $\underset{\sim}{\lambda}_{(i)}$ denote the ith column of $\underset{\sim}{\Lambda}$, then

(5.2.2) $$\underset{\sim}{\lambda}'_{(i)}\ \underset{\sim}{\lambda}_{(j)} = \lambda_{ij}.$$

Proof : We have

$$\underset{\sim}{\Lambda}'\ \underset{\sim}{\Lambda} = \begin{bmatrix} \underset{\sim}{\lambda}'_{(1)} \\ \underset{\sim}{\lambda}'_{(2)} \\ \vdots \\ \underset{\sim}{\lambda}'_{(n)} \end{bmatrix} [\underset{\sim}{\lambda}_{(1)}\ \underset{\sim}{\lambda}_{(2)} \cdots \underset{\sim}{\lambda}_{(n)}] = ((\underset{\sim}{\lambda}'_{(i)}\ \underset{\sim}{\lambda}_{(j)})).$$

Since $\underset{\sim}{\Lambda}$ is symmetric and idempotent, hence $\underset{\sim}{\Lambda}'\ \underset{\sim}{\Lambda} = \underset{\sim}{\Lambda}$, and the result follows on equating the (i,j)th element on both sides.

Property (iii) : $\underset{\sim}{\lambda}'_{(i)}\ \underset{\sim}{X} = \underset{\sim}{0}'$.

This is obvious, since $\underset{\sim}{\Lambda}\underset{\sim}{X} = \underset{\sim}{0}$. The distribution of u_{ij} under H_{12} is given in Theorem 5.2.1. Its proof requires the following lemma.

Lemma 5.2.1. Let

(5.2.3) $$t_{ij} = (e_i/\lambda_{ii}^{1/2} + e_j/\lambda_{jj}^{1/2})/[2(1+\rho_{ij})]^{1/2},$$

$$Q_1 = t_{ij}^2 \text{ and } Q_2 = S^2 - t_{ij}^2.$$ Then under H_{12},

(i) t_{ij} is distributed as $N(\mu, \sigma^2)$, where

(5.2.4) $$\mu = [(\rho_{1i}+\rho_{1j})\ \lambda_{11}^{1/2}\ \theta_1 + (\rho_{2i}+\rho_{2j})\lambda_{22}^{1/2}\ \theta_2]/[2(1+\rho_{ij})]^{1/2}.$$

This will be denoted by $t_{ij} \overset{d}{=} N(\mu, \sigma^2)$.

(ii) $Q_1 \overset{d}{=} \sigma^2\chi^2(1, \delta)$ and $Q_2 \overset{d}{=} \sigma^2\chi^2(n-k-1, \eta)$,

where $\chi^2(a,\lambda)$ denotes a non-central χ^2 distribution with 'a' degrees of freedom and non-centrality parameter λ, $\sigma^2\delta = \mu^2$ and

$$(5.2.5)\quad \sigma^2\eta = [1/\{2(1+\rho_{ij})\}]\,[\{2(1+\rho_{ij})-(\rho_{1i}+\rho_{1j})^2\}\,\lambda_{11}\,\theta_1^2 + \{2(1+\rho_{ij})-(\rho_{2i}+\rho_{2j})^2\}\,\lambda_{22}\,\theta_2^2 + 2\{2\lambda_{12}(1+\rho_{ij})-(\rho_{1i}+\rho_{1j})(\rho_{2i}+\rho_{2j})(\lambda_{11}\lambda_{22})^{\frac{1}{2}}\}\theta_1\theta_2]$$

(iii) t_{ij} and Q_2 are independent.

Proof : Since $e_i = \underset{\sim}{\lambda}'_{(i)}\,\underset{\sim}{y}$, hence from equation (5.2.3) we have

$$t_{ij} = \frac{1}{[2(1+\rho_{ij})]^{1/2}}\left[\frac{\underset{\sim}{\lambda}'_{(i)}}{\lambda_{ii}^{1/2}} + \frac{\underset{\sim}{\lambda}'_{(j)}}{\lambda_{jj}^{1/2}}\right]\underset{\sim}{y} = \underset{\sim}{c}'\,\underset{\sim}{y},$$

where

$$\underset{\sim}{c}' = \frac{1}{[2(1+\rho_{ij})]^{1/2}}\left[\frac{\underset{\sim}{\lambda}'_{(i)}}{\lambda_{ii}^{1/2}} + \frac{\underset{\sim}{\lambda}'_{(j)}}{\lambda_{jj}^{1/2}}\right].$$

Note that

$$\underset{\sim}{c}'\underset{\sim}{c} = \frac{1}{2(1+\rho_{ij})}\left[\frac{\underset{\sim}{\lambda}'_{(i)}\underset{\sim}{\lambda}_{(i)}}{\lambda_{ii}} + \frac{2\underset{\sim}{\lambda}'_{(i)}\underset{\sim}{\lambda}_{(j)}}{(\lambda_{ii}\,\lambda_{jj})^{1/2}} + \frac{\underset{\sim}{\lambda}'_{(j)}\underset{\sim}{\lambda}_{(j)}}{\lambda_{jj}}\right]$$

$$= \frac{1}{2(1+\rho_{ij})}\,(2 + 2\rho_{ij}),$$

on using equation (5.2.2). Consequently,

$$(5.2.6)\qquad \underset{\sim}{c}'\,\underset{\sim}{c} = 1.$$

Since t_{ij} is a linear combination of $\underset{\sim}{y}$ and under H_{12}

$$\underset{\sim}{y} \overset{d}{=} N(\underset{\sim}{X}\,\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}_1\,\theta_1 + \underset{\sim}{\varepsilon}_2\,\theta_2,\ \sigma^2\,\underset{\sim}{I}),$$

hence under H_{12}, t_{ij} is normally distributed with mean and variance, respectively, given by

$$\mu = E(t_{ij}) = \underset{\sim}{c}' E(\underset{\sim}{y}) = \underset{\sim}{c}' (X \underset{\sim}{\beta} + \underset{\sim}{\varepsilon}_1 \theta_1 + \underset{\sim}{\varepsilon}_2 \theta_2)$$

$$= \underset{\sim}{c}'(\underset{\sim}{\varepsilon}_1 \theta_1 + \underset{\sim}{\varepsilon}_2 \theta_2)$$

$$= \frac{1}{[2(1+\rho_{ij})]^{1/2}} \left[\frac{\underset{\sim}{\lambda}'(i)\underset{\sim}{\varepsilon}_1\theta_1}{\lambda_{ii}^{1/2}} + \frac{\underset{\sim}{\lambda}'(i)\underset{\sim}{\varepsilon}_2\theta_2}{\lambda_{ii}^{1/2}} + \frac{\underset{\sim}{\lambda}'(j)\underset{\sim}{\varepsilon}_1\theta_1}{\lambda_{jj}^{1/2}} + \frac{\underset{\sim}{\lambda}'(j)\underset{\sim}{\varepsilon}_2\theta_2}{\lambda_{jj}^{1/2}}\right]$$

$$= \frac{1}{[2(1+\rho_{ij})]^{1/2}} \left[\left(\frac{\lambda_{1i}}{\lambda_{ii}^{1/2}} + \frac{\lambda_{1j}}{\lambda_{jj}^{1/2}}\right) \theta_1 + \left(\frac{\lambda_{2i}}{\lambda_{ii}^{1/2}} + \frac{\lambda_{2j}}{\lambda_{jj}^{1/2}}\right) \theta_2\right]$$

$$= [(\rho_{1i}+\rho_{1j})\, \lambda_{11}^{1/2}\theta_1 + (\rho_{2i}+\rho_{2j})\, \lambda_{22}^{1/2}\theta_2]/[2(1+\rho_{ij})]^{1/2},$$

$$\text{Var}(t_{ij}) = \sigma^2\, \underset{\sim}{c}'\, \underset{\sim}{c} = \sigma^2,$$

on using equation (5.2.6).

Next, $Q_1 = t_{ij}^2 = \underset{\sim}{y}'\, \underset{\sim}{c}\, \underset{\sim}{c}'\, \underset{\sim}{y}$.

Hence $Q_1 \overset{d}{=} \sigma^2 \chi^2(1, \delta)$, where the non-centrality parameter is given by

$$\sigma^2 \delta = [E(\underset{\sim}{y})]'\, \underset{\sim}{c}\, \underset{\sim}{c}'\, [E(\underset{\sim}{y})]$$

$$= [E(t_{ij})]^2 = \mu^2.$$

Further $Q_2 = S^2 - t_{ij}^2$

$$= \underset{\sim}{y}'\, \underset{\sim}{\Lambda}\, \underset{\sim}{y} - \underset{\sim}{y}'\, \underset{\sim}{c}\, \underset{\sim}{c}'\, \underset{\sim}{y}$$

$$= \underset{\sim}{y}'\, [\underset{\sim}{\Lambda} - \underset{\sim}{c}\, \underset{\sim}{c}']\, \underset{\sim}{y}.$$

Thus Q_2 is also a quadratic form in $\underset{\sim}{y}$. The matrix of the quadratic form is $\underset{\sim}{\Lambda} - \underset{\sim}{c}\, \underset{\sim}{c}'$, which satisfies

$$(\underset{2}{\Lambda} - \underset{2}{c}\,\underset{2}{c}')^2 = (\underset{2}{\Lambda} - \underset{2}{c}\,\underset{2}{c}')\,(\underset{2}{\Lambda} - \underset{2}{c}\,\underset{2}{c}')$$

$$= \underset{2}{\Lambda}^2 - \underset{2}{\Lambda}\,\underset{2}{c}\,\underset{2}{c}' - \underset{2}{c}\,\underset{2}{c}'\,\underset{2}{\Lambda} + \underset{2}{c}\,\underset{2}{c}'\,\underset{2}{c}\,\underset{2}{c}'$$

$$= \underset{2}{\Lambda} - \underset{2}{\Lambda}\,\underset{2}{c}\,\underset{2}{c}' - \underset{2}{c}\,\underset{2}{c}'\,\underset{2}{\Lambda} + \underset{2}{c}\,\underset{2}{c}'$$

on using equation (5.2.6). Further,

$$\underset{2}{\Lambda}\,\underset{2}{c} = \begin{bmatrix} \underset{2}{\lambda}'(1) \\ \underset{2}{\lambda}'(2) \\ \vdots \\ \underset{2}{\lambda}'(n) \end{bmatrix} \frac{1}{[2(1+\rho_{ij})]^{1/2}} \left[\frac{\underset{2}{\lambda}(i)}{\lambda_{ii}^{1/2}} + \frac{\underset{2}{\lambda}(j)}{\lambda_{jj}^{1/2}}\right]$$

$$= \frac{1}{[2(1+\rho_{ij})]^{1/2}} \begin{bmatrix} \lambda_{1i}/\lambda_{ii}^{1/2} + \lambda_{1j}/\lambda_{jj}^{1/2} \\ \lambda_{2i}/\lambda_{ii}^{1/2} + \lambda_{2j}/\lambda_{jj}^{1/2} \\ \vdots \\ \lambda_{ni}/\lambda_{ii}^{1/2} + \lambda_{nj}/\lambda_{jj}^{1/2} \end{bmatrix}$$

$$= \frac{1}{[2(1+\rho_{ij})]^{1/2}} \left[\frac{\underset{2}{\lambda}(i)}{\lambda_{ii}^{1/2}} + \frac{\underset{2}{\lambda}(j)}{\lambda_{jj}^{1/2}}\right]$$

$$= \underset{2}{c}.$$

This implies that $\underset{2}{c}'\,\underset{2}{\Lambda} = \underset{2}{c}'$ and $(\underset{2}{\Lambda} - \underset{2}{c}\,\underset{2}{c}')^2 = \underset{2}{\Lambda} - \underset{2}{c}\,\underset{2}{c}'$. Thus $\underset{2}{\Lambda} - \underset{2}{c}\,\underset{2}{c}'$ is an idempotent matrix of rank given by

$$\text{rank}\,(\underset{2}{\Lambda} - \underset{2}{c}\,\underset{2}{c}') = \text{tr}\,(\underset{2}{\Lambda} - \underset{2}{c}\,\underset{2}{c}')$$

$$= \text{tr}\,\underset{2}{\Lambda} - \text{tr}\,(\underset{2}{c}\,\underset{2}{c}')$$

$$= \text{rank}\,(\Lambda) - \text{tr}\,(\underset{2}{c}'\,\underset{2}{c})$$

$$= (n-k) - \underset{2}{c}'\,\underset{2}{c}$$

$$= n-k-1.$$

Consequently,

$$Q_2 \overset{d}{=} \sigma^2 \chi^2(n-k-1, \eta),$$

where

$$\begin{aligned}\sigma^2\eta &= [E(\underset{\sim}{y})]' \, [\Lambda - \underset{\sim}{c}\,\underset{\sim}{c}'] \, [E(\underset{\sim}{y})] \\ &= E(\underset{\sim}{y}') \, \Lambda \, E(\underset{\sim}{y}) - E(\underset{\sim}{y}') \, \underset{\sim}{c}\,\underset{\sim}{c}' \, E(\underset{\sim}{y}) \\ &= \sigma^2\lambda^* - \sigma^2\delta,\end{aligned}$$

where λ^* is given at equation (5.2.1) with $i = 1$ and $j = 2$, and $\sigma^2\delta = \mu^2$. Hence

$$\begin{aligned}\sigma^2\eta &= \lambda_{11}\theta_1^2 + \lambda_{22}\theta_2^2 + 2\lambda_{12}\,\theta_1\theta_2 \\ &\quad - [1/\{2(1+\rho_{ij})\}] \, [(\rho_{1i}+\rho_{1j})^2 \, \lambda_{11}\theta_1^2 + (\rho_{2i}+\rho_{2j}) \, \lambda_{22}\theta_2^2 \\ &\qquad + 2(\rho_{1i}+\rho_{1j})(\rho_{2i}+\rho_{2j})(\lambda_{11}\lambda_{22})^{1/2} \, \theta_1\theta_2] \\ &= [1/\{2(1+\rho_{ij})\}] \, [\{2(1+\rho_{ij}) - (\rho_{1i}+\rho_{1j})^2\}\lambda_{11}\theta_1^2 \\ &\qquad + \{2(1+\rho_{ij}) - (\rho_{2i}+\rho_{2j})^2\} \, \lambda_{22}\theta_2^2 \\ &\qquad + 2\{2\lambda_{12}(1+\rho_{ij}) - (\rho_{1i}+\rho_{1j})(\rho_{2i}+\rho_{2j})(\lambda_{11}\lambda_{22})^{1/2}\} \, \theta_1\theta_2] .\end{aligned}$$

Finally, $t_{ij} = \underset{\sim}{c}' \, \underset{\sim}{y}$ and $Q_2 = \underset{\sim}{y}'(\Lambda - \underset{\sim}{c}\,\underset{\sim}{c}')\underset{\sim}{y}$ are independent, since

$$(\Lambda - \underset{\sim}{c}\,\underset{\sim}{c}')\underset{\sim}{c} = \Lambda\,\underset{\sim}{c} - \underset{\sim}{c}\,\underset{\sim}{c}'\,\underset{\sim}{c} = \underset{\sim}{c} - \underset{\sim}{c} = \underset{\sim}{0}.$$

This completes the proof of the lemma.

We next obtain a general distribution from which the distribution of u_{ij} under H_{12} can be obtained immediately.

Theorem 5.2.1. Let $T \overset{d}{=} N(\mu,1)$ and $Q \overset{d}{=} \chi^2(a,\eta)$. If T and Q are independently distributed, then

$$Z = T/(T^2 + Q)^{1/2} \tag{5.2.7}$$

has a pdf given by

$$f(z) = \sum_{j=0}^{\infty} K_j \sum_{i=0}^{\infty} K_i^* \frac{z^i(1 - z^2)^{j+a/2-1}}{B[(i+1)/2,\ j+a/2]}, \quad -1 \le z \le 1, \tag{5.2.8}$$

where

$$K_j = e^{-\eta/2} (\tfrac{\eta}{2})^j/j!\ , \quad j = 0,1,2,\ldots, \text{ and} \tag{5.2.9}$$

$$K_i^* = e^{-\mu^2/2}\, \mu^i 2^{-i/2}/G(i/2+1), \quad i = 0,1,2,\ldots\ . \tag{5.2.10}$$

Proof : Since $T \overset{d}{=} N(\mu,1)$, $Q \overset{d}{=} \chi^2(a,\eta)$ and T and Q are independent, hence the joint pdf of T and Q is

$$f(t,Q) = \frac{1}{(2\pi)^{1/2}} e^{-(t-\mu)^2/2} \cdot e^{-\eta/2} \sum_{j=0}^{\infty} \frac{1}{j!\ 2^{j+a/2}\ G(j+a/2)} (\tfrac{\eta}{2})^j\ e^{-Q/2}\ Q^{j+a/2-1},$$

$$0 < Q < \infty, \quad -\infty < t < \infty$$

$$= \sum_{j=0}^{\infty} K_j \frac{e^{-\mu^2/2}}{(2\pi)^{1/2}\ 2^{j+a/2}\ G(j+a/2)} e^{-(t^2-2\mu t+Q)/2} Q^{j+a/2-1},$$

where K_j is given at equation (5.2.9).

Now making a transformation

$$z = t/(t^2+Q)^{1/2}, \text{ and}$$

$$Q = Q, \quad -1 \le z \le 1,$$

the inverse transformation is

$$t = z\, Q^{1/2}/(1-z^2)^{1/2},$$

$$Q = Q.$$

The jacobian of transformation is given by

$$\left|\frac{\partial(t,Q)}{\partial(z,Q)}\right| = \begin{vmatrix} Q^{1/2}/(1-z^2)^{3/2} & z/[2\{Q(1-z^2)\}^{1/2}] \\ 0 & 1 \end{vmatrix}$$

$$= Q^{1/2}/(1-z^2)^{3/2},$$

and the joint pdf of Z and Q is given by

$$(5.2.11)\quad f(z,Q) = \sum_{j=0}^{\infty} K_j \frac{e^{-\mu^2/2}}{(2\pi)^{1/2}\, 2^{j+a/2}\, G(j+a/2)}$$

$$\cdot\, e^{-\{z^2Q/(1-z^2)-2\mu\, zQ^{1/2}/(1-z^2)^{1/2} + Q\}/2}$$

$$\cdot\, Q^{j+(a-1)/2}\, (1-z^2)^{-3/2}$$

$$= \sum_{j=0}^{\infty} K_j \frac{e^{-\mu^2/2}}{(2\pi)^{1/2}\, 2^{j+a/2}\, G(j+a/2)}$$

$$\cdot\, e^{-\{Q/(1-z^2)-2\mu z Q^{1/2}/(1-z^2)^{1/2}\}/2}\, Q^{j+(a-1)/2}(1-z^2)^{-3/2}$$

$$0 < Q < \infty,\ -1 \le z \le 1.$$

Make a transformation $x = \dfrac{Q^{1/2}}{(1-z^2)^{1/2}}$, $z = z$; that is,

$$Q = x^2(1-z^2),\ z = z.$$

The jacobian of transformation is $\left|\dfrac{\partial(Q,z)}{\partial(x,z)}\right| = 2x(1-z^2)$,

and the joint distribution of Z and X becomes

$$f(z,x) = \sum_{j=0}^{\infty} K_j \frac{e^{-\mu^2/2}}{(2\pi)^{1/2}\, 2^{j+a/2}\, G(j+a/2)}$$

$$\cdot\, e^{-(x^2-2\mu zx)/2}\, x^{2j+a-1}\, (1-z^2)^{j+(a-1)/2} (1-z^2)^{-3/2}\, 2x(1-z^2)$$

$$= \sum_{j=0}^{\infty} K_j \frac{e^{-\mu^2/2}}{(2\pi)^{1/2}\, 2^{j+a/2-1}\, G(j+a/2)} (1-z^2)^{j+a/2-1}$$

$$\cdot\, x^{2j+a}\, e^{-(x^2-2\mu zx)/2}\,, \quad -1 \le z \le 1,\ 0 \le x \le \infty.$$

Integrating x from 0 to ∞,

$$(5.2.12)\quad f(z) = \sum_{j=0}^{\infty} K_j \frac{e^{-\mu^2/2}\, (1-z^2)^{j+a/2-1}}{(2\pi)^{1/2}\, 2^{j+a/2-1}\, G(j+a/2)}$$

$$\cdot \int_0^{\infty} e^{-x^2/2}\, x^{2j+a}\, e^{\mu zx}\, dx.$$

Now, consider the integral

$I(z) = \int_0^{\infty} e^{-x^2/2}\, x^{2j+a}\, e^{\mu xz}\, dx$ appearing in equation (5.2.12). This integral is convergent. Expanding $\exp(\mu xz)$ in powers of μxz, we get

$$I(z) = \int_0^{\infty} \sum_{i=0}^{\infty} \{(\mu zx)^i/i!\}\, x^{2j+a}\, e^{-x^2/2}\, dx$$

$$= \sum_{i=0}^{\infty} (\mu^i z^i/i!) \int_0^{\infty} x^{2\{j+a/2+(i+1)/2\}-1}\, e^{-x^2/2}\, dx,\ -1 \le z \le 1.$$

Letting

$y = x^2/2$, that is $x = (2y)^{1/2}$, we have

$xdx = dy$, and

$$I(z) = 2^{j+a/2+(i+1)/2-1} \sum_{i=0}^{\infty} (\mu^i z^i/i!) \int_0^{\infty} y^{j+a/2+(i+1)/2-1} . e^{-y}\, dy$$

$$= 2^{j+a/2+(i+1)/2-1} \sum_{i=0}^{\infty} (\mu^i z^i/i!)\, G\{j+a/2+(i+1)/2\}.$$

Substituting for $I(z)$ in equation (5.2.12) we get

$$f(z) = \sum_{j=0}^{\infty} K_j \frac{e^{-\mu^2/2}(1-z^2)^{j+a/2-1}\, 2^{j+a/2+i/2-1/2}}{\pi^{1/2}\, 2^{j+a/2-1/2}\, G(j+a/2)}$$

$$\cdot \sum_{i=0}^{\infty} (\mu^i z^i/i!)\, G\{j+a/2+(i+1)/2\}$$

$$= \sum_{j=0}^{\infty} K_j \sum_{i=0}^{\infty} \frac{e^{-\mu^2/2}\, 2^{i/2}\, G\{j+a/2+(i+1)/2\}\, G\{(i+1)/2\}}{\pi^{1/2}\, i!\, G(j+a/2)\, G\{(i+1)/2\}}$$

$$\cdot \mu^i z^i (1-z^2)^{j+a/2-1}$$

$$= \sum_{j=0}^{\infty} K_j \sum_{i=0}^{\infty} \frac{e^{-\mu^2/2}\, 2^{i/2}\, G\{(i+1)/2\}\, \mu^i}{i!\, G(1/2)\, B[(i+1)/2, j+a/2]}\, z^i (1-z^2)^{j+a/2-1}$$

$$= \sum_{j=0}^{\infty} K_j \sum_{i=0}^{\infty} K_i^{*}\, z^i (1-z^2)^{j+a/2-1}/B[(i+1)/2,\ j+a/2], \quad -1 \le z \le 1,$$

where

$$K_i^{*} = \frac{e^{-\mu^2/2}\, 2^{i/2}\, G\{(i+1)/2\}\, \mu^i}{i!\, G(1/2)}, \quad i = 0,1,2,\ldots .$$

Next, writing $i! = G(i+1)$ and applying the duplication formula

$$G(2x) = (2\pi)^{-1/2}\, 2^{2x-1/2}\, G(x)\, G(x+1/2),$$

we get

$$i! = \pi^{-1/2}\, 2^{i}\, G\{(i+1)/2\}\; G(i/2+1)\,.$$

Consequently,

$$K_i^* = \frac{e^{-\mu^2/2}\;\mu^{i}\;2^{-i/2}\;G\{(i+1)/2\}}{G(i/2+1)\;G\{(i+1)/2\}}$$

$$= e^{-\mu^2/2}\;\mu^{i}\;2^{-i/2}/G(i/2+1)\,.$$

This completes the proof of the theorem.

Corollary 5.2.1. Let $Z_1 = Z^2$. Then the pdf of Z_1 is given by

$$(5.2.13)\quad f(z_1) = \sum_{j=0}^{\infty} K_j \sum_{i=0}^{\infty} K_{2i}^* \frac{z_1^{i-1/2}(1-z_1)^{j+a/2-1}}{B(i+1/2, j+a/2)}; 0 \le z_1 \le 1.$$

Proof : From Theorem 5.2.1, we have

$$f(z) = \sum_{j=0}^{\infty} K_j \sum_{i_1=0}^{\infty} K_{i_1}^* z^{i_1}(1-z^2)^{j+a/2-1}/B[(i_1+1)/2, j+a/2]\,.$$

Making a transformation $z_1 = z^2$, the inverse transformations are

$$z = z_1^{1/2} \text{ and } z = -z_1^{1/2}.$$

Hence the jacobian of transformation $\left|\frac{\partial z}{\partial z_1}\right|$ is $1/(2z_1^{1/2})$ for both the cases. Consequently,

$$f(z_1) = f(z)/(2z_1^{1/2})\Big|_{z=z_1^{1/2}} + f(z)/(2z_1^{1/2})\Big|_{z=-z_1^{1/2}}$$

$$= 2 \sum_{j=0}^{\infty} K_j \sum_{i_1=0}^{\infty} K_{i_1}^* (1/2)$$

$$\cdot\, z_1^{(i_1-1)/2}\,(1-z_1)^{j+a/2-1}/B[(i_1+1)/2, j+a/2]; (i_1 \text{ is even})$$

$$= \sum_{j=0}^{\infty} K_j \sum_{\substack{i_1=0 \\ (i_1 \text{ even})}}^{\infty} K_{i_1}^{*} z_1^{(i_1+1)/2-1} (1-z_1)^{j+a/2-1}/B[(i_1+1)/2, j+a/2].$$

Writing $i_1 = 2i$, we simply get

$$f(z_1) = \sum_{j=0}^{\infty} K_j \sum_{i=0}^{\infty} K_{2i}^{*} z_1^{i-1/2}(1-z_1)^{j+a/2-1}/B(i+1/2, j+a/2).$$

Note that

$$K_{2i}^{*} = e^{-\mu^2/2} (\mu^2/2)^i/G(i+1), \quad i = 0,1,2,\ldots$$

$$= e^{-\mu^2/2} (\mu^2/2)^i/i!\,.$$

Thus the marginal distribution of Z_1 is a linear combination of beta variables with weights given by Poisson probability terms.

Now, we apply the results of Theorem 5.2.1 to get the pdf of u_{ij}.

From equations(5.1.1) and (5.2.3), we have

$$u_{ij} = t_{ij}/S = t_{ij}/(t_{ij}^2 + Q_2)^{1/2}$$

$$= \frac{t_{ij}/\sigma}{(t_{ij}^2/\sigma^2 + Q_2/\sigma^2)^{1/2}}\,.$$

By Lemma 5.2.1, we see that under H_{12}, $t_{ij}/\sigma \overset{d}{=} N(\mu/\sigma, 1)$ and $Q_2 \overset{d}{=} \sigma^2\chi^2(n-k-1, \eta)$, μ and η are defined by equations (5.2.4) and (5.2.5) respectively.

Further t_{ij} and Q_2 are independent. The conditions of Theorem 5.2.1 are thus satisfied. Hence we have the following

theorem for $\mu > 0$ (with obvious changes for $\mu < 0$).

Theorem 5.2.2. The non-null distribution of u_{ij} under H_{12} is given by

(5.2.14) $f(u_{ij})$

$$= \sum_{j_1=0}^{\infty} K_{j_1} \sum_{i_1=0}^{\infty} K^*_{i_1} (u_{ij})^{i_1} (1-u_{ij}^2)^{j_1+a/2-1} / B[(i_1+1)/2, j_1+a/2],$$

$$-1 \leq u_{ij} \leq 1,$$

where (for $\mu > 0$)

$$K_{j_1} = e^{-\eta/2} (\eta/2)^{j_1}/j_1! \quad ; \; j_1 = 0,1,2,\ldots,$$

$$K^*_{i_1} = e^{-\delta/2} (\delta/2)^{i_1/2}/G(i_1/2 + 1); \; i_1 = 0,1,2,\ldots,$$

$a = n-k-1$, $\delta = \mu^2/\sigma^2$; μ and η are as defined in equations (5.2.4) and (5.2.5) respectively.

Note that $\theta_1 = \theta_2 = 0$ implies $\eta = \mu = 0$ and $K_o = K^*_o = 1$, while other K_{j_1} and $K^*_{i_1}$ are zero. Also $a = n-k-1 = p-1$. From equation (5.2.14) we immediately get the pdf of u_{ij} under the null hypothesis as

$$f(u_{ij}) = (1-u_{ij}^2)^{(p-3)/2}/B[1/2,(p-1)/2] \; , \; -1 \leq u_{ij} \leq 1,$$

which is same as the null distribution obtained in equation (2.2.11).

For studying the performance of our statistic, we need

$$p^{i,j} = \Pr (u_{ij} > u_\alpha | H_{12}).$$

The probability $P^{i,j}$ can be obtained from Corollary 5.2.2.

Corollary 5.2.2. For $u_\alpha \geq 0$,

$$(5.2.15)\quad \Pr(u_{ij} > u_\alpha | H_{12}) = \frac{1}{2} \sum_{j=0}^{\infty} K_j \sum_{i=0}^{\infty} K_i^* I_{1-u_\alpha^2}[j+a/2, (i+1)/2],$$

where i and j are used for i_1 and j_1 respectively and u is used for variable of integration for convenience.

Proof : $\Pr(u_{ij} > u_\alpha | H_{12})$

$$= \sum_{j=0}^{\infty} K_j \sum_{i=0}^{\infty} K_i^* \int_{u_\alpha}^{1} u^i (1-u^2)^{j+a/2-1} du / B[(i+1)/2, j+a/2].$$

Putting $y = u^2$, $dy = 2u\, du$, we get

$\Pr(u_{ij} > u_\alpha | H_{12})$

$$= \frac{1}{2} \sum_{j=0}^{\infty} K_j \sum_{i=0}^{\infty} K_i^* \frac{1}{B[(i+1)/2, j+a/2]} \int_{u_\alpha^2}^{1} y^{(i+1)/2-1} (1-y)^{j+a/2-1} dy$$

$$= \frac{1}{2} \sum_{j=0}^{\infty} K_j \sum_{i=0}^{\infty} K_i^* I_{1-u_\alpha^2}[j+a/2, (i+1)/2],$$

on substituting $z = 1-y$ in the integral.

This proves the corollary.

Note that

$$\Pr(u_{ij} > 0 | H_{12}) = \Pr(t_{ij} > 0 | H_{12}) = \Phi(\mu/\sigma).$$

Consequently, this probability is an increasing function of μ. Thus for large values of μ/σ, the distribution of u_{ij} is essentially confined to the interval $[0,1]$, for example, for $\mu/\sigma = 3$, we have $Pr(u_{ij} > 0|H_{12}) = 0.99865$. This is useful for evaluating an approximate expression for $Pr(u_{ij} > u_\alpha|H_{12})$ for large values of μ/σ.

5.3. Non-null distribution of v_{ij}

The statistic V is useful, when there are two outliers, one on either extreme. So we consider the alternative hypothesis model under this set up with one observation having a mean greater than the one under H_o and another observation having a mean less than the one specified under H_o. Thus under H^*_{ij}, $i \neq j$,

$$E(\underset{\sim}{y}) = \underset{\sim}{X}\,\underset{\sim}{\beta} - \underset{\sim}{\varepsilon}_i\,\theta_i + \underset{\sim}{\varepsilon}_j\,\theta_j,$$

where $\underset{\sim}{\varepsilon}_i$ $(i = 1,2,\ldots,n)$ is the ith column of the identity matrix $\underset{\sim}{I}_n$ and θ_i, $\theta_j > 0$.

In this case the alternative hypothesis is the union of $n(n-1)$ hypotheses $\bigcup_{\substack{i,j=1,2,\ldots,n \\ i\neq j}} H^*_{ij}$. Under H^*_{ij}, $\underset{\sim}{e}$ is normally distributed with variance-covariance matrix $\underset{\sim}{\Lambda}\sigma^2$ and mean

$$\begin{aligned} E(\underset{\sim}{e}|H^*_{ij}) &= \underset{\sim}{\Lambda}\, E(\underset{\sim}{y}|H^*_{ij}) \\ &= \underset{\sim}{\Lambda}\,(\underset{\sim}{X}\,\underset{\sim}{\beta} - \underset{\sim}{\varepsilon}_i\,\theta_i + \underset{\sim}{\varepsilon}_j\,\theta_j) \\ &= \underset{\sim}{\Lambda}\,(\underset{\sim}{\varepsilon}_j\,\theta_j - \underset{\sim}{\varepsilon}_i\,\theta_i) \end{aligned}$$

since $\underset{\sim}{\Lambda}\underset{\sim}{X} = \underset{\sim}{0}$.

The residual sum of squares S^2 has a non-central $\sigma^2\chi^2$ distribution with n-k degrees of freedom and non-centrality parameter λ^{**} , where

$$\sigma^2 \lambda^{**} = E(\underset{\sim}{y}' | H^*_{ij}) \underset{\sim}{\Lambda} E(\underset{\sim}{y} | H^*_{ij})$$

$$= (\underset{\sim}{\beta}' \underset{\sim}{X}' - \underset{\sim}{\varepsilon}'_i \theta_i + \underset{\sim}{\varepsilon}'_j \theta_j) \underset{\sim}{\Lambda} (\underset{\sim}{X} \underset{\sim}{\beta} - \underset{\sim}{\varepsilon}_i \theta_i + \underset{\sim}{\varepsilon}_j \theta_j)$$

$$(5.3.1) = \lambda_{ii} \theta_i^2 + \lambda_{jj} \theta_j^2 - 2 \lambda_{ij} \theta_i \theta_j .$$

Again for the sake of convenience we consider H^*_{12} and derive the exact distribution of v_{ij} under H^*_{12}. For this we use the properties of $\underset{\sim}{\Lambda}$ mentioned in Section 5.2. Its proof requires the following lemma.

<u>Lemma 5.3.1</u>. Let

$$(5.3.2) \quad t^*_{ij} = (e_i/\lambda_{ii}^{1/2} - e_j/\lambda_{jj}^{1/2})/[2(1-\rho_{ij})]^{1/2},$$

$Q_1 = t^{*2}_{ij}$ and $Q_2 = S^2 - t^{*2}_{ij}$. Then under H^*_{12},

(i) t^*_{ij} is distributed as $N(\mu^*, \sigma^2)$, where

$$(5.3.3) \quad \mu^* = [(\rho_{1j}-\rho_{1i}) \lambda_{11}^{1/2} \theta_1 - (\rho_{2j}-\rho_{2i}) \lambda_{22}^{1/2} \theta_2]/[2(1-\rho_{ij})]^{1/2}$$

(ii) $Q_1 \overset{d}{=} \sigma^2\chi^2(1, \delta^*)$ and $Q_2 \overset{d}{=} \sigma^2\chi^2(n-k-1, \eta^*)$,

where $\chi^2(a, \lambda)$ denotes a non-central χ^2 distribution with 'a' degrees of freedom and noncentrality parameter λ , $\sigma^2 \delta^* = \mu^{*2}$ and

$$(5.3.4)\quad \sigma^2 \eta^* = [1/\{2(1-\rho_{ij})\}]\,[\{2(1-\rho_{ij})-(\rho_{1j}-\rho_{1i})^2\}\,\lambda_{11}\theta_1^2$$

$$+\{2(1-\rho_{ij})-(\rho_{2j}-\rho_{2i})^2\}\,\lambda_{22}\theta_2^2-2\{2\lambda_{12}(1-\rho_{ij})$$

$$-(\rho_{1j}-\rho_{1i})(\rho_{2j}-\rho_{2i})(\lambda_{11}\lambda_{22})^{1/2}\}\,\theta_1\theta_2\,]$$

(iii) t^*_{ij} and Q_2 are independent, and hence Q_1 and Q_2 independent.

<u>Proof</u> : Since $e_i = \underset{\sim}{\lambda}'_{(i)}\,\underset{\sim}{y}$, hence from equation (5.3.2), we have

$$t^*_{ij} = \frac{1}{[2(1-\rho_{ij})]^{1/2}}\left[\frac{\underset{\sim}{\lambda}'_{(i)}}{\lambda_{ii}^{1/2}} - \frac{\underset{\sim}{\lambda}'_{(j)}}{\lambda_{jj}^{1/2}}\right]\underset{\sim}{y} = \underset{\sim}{d}'\,\underset{\sim}{y},$$

where

$$\underset{\sim}{d}' = \frac{1}{[2(1-\rho_{ij})]^{1/2}}\left[\frac{\underset{\sim}{\lambda}'_{(i)}}{\lambda_{ii}^{1/2}} - \frac{\underset{\sim}{\lambda}'_{(j)}}{\lambda_{jj}^{1/2}}\right].$$

Note that

$$\underset{\sim}{d}'\underset{\sim}{d} = \frac{1}{2(1-\rho_{ij})}\left[\frac{\underset{\sim}{\lambda}'_{(i)}\,\underset{\sim}{\lambda}_{(i)}}{\lambda_{ii}} - \frac{2\underset{\sim}{\lambda}'_{(i)}\,\underset{\sim}{\lambda}_{(j)}}{(\lambda_{ii}\,\lambda_{jj})^{1/2}} + \frac{\underset{\sim}{\lambda}'_{(j)}\,\underset{\sim}{\lambda}_{(j)}}{\lambda_{jj}}\right]$$

$$= \frac{1}{2(1-\rho_{ij})}\,(2 - 2\rho_{ij}),$$

on using equation (5.2.2). Consequently,

$$(5.3.5)\qquad \underset{\sim}{d}'\,\underset{\sim}{d} = 1.$$

Since t^*_{ij} is a linear combination of $\underset{\sim}{y}$ and under H^*_{12}

$$\underset{\sim}{y} \overset{d}{=} N(X\,\underset{\sim}{\beta} - \underset{\sim}{\varepsilon}_1\theta_1 + \underset{\sim}{\varepsilon}_2\theta_2, \sigma^2\,\underset{\sim}{I}),$$

hence t^*_{ij} is normally distributed under H^*_{12} with mean and variance given by

$$\mu^* = E(t_{ij}^*) = \underset{\sim}{d}' E(\underset{\sim}{y}) = \underset{\sim}{d}'(\underset{\sim}{X}\,\underset{\sim}{\beta} - \underset{\sim}{\varepsilon}_1\theta_1 + \underset{\sim}{\varepsilon}_2\theta_2)$$

$$= \underset{\sim}{d}'(\underset{\sim}{\varepsilon}_2\theta_2 - \underset{\sim}{\varepsilon}_1\theta_1)$$

$$= \frac{1}{[2(1-\rho_{ij})]^{1/2}}\left[\frac{\underset{\sim}{\lambda}'(i)\underset{\sim}{\varepsilon}_2\theta_2}{\lambda_{ii}^{1/2}} - \frac{\underset{\sim}{\lambda}'(i)\underset{\sim}{\varepsilon}_1\theta_1}{\lambda_{ii}^{1/2}} - \frac{\underset{\sim}{\lambda}'(j)\underset{\sim}{\varepsilon}_2\theta_2}{\lambda_{jj}^{1/2}} + \frac{\underset{\sim}{\lambda}'(j)\underset{\sim}{\varepsilon}_1\theta_1}{\lambda_{jj}^{1/2}}\right]$$

$$= \frac{1}{[2(1-\rho_{ij})]^{1/2}}\left[\left(\frac{\lambda_{1j}}{\lambda_{jj}^{1/2}} - \frac{\lambda_{1i}}{\lambda_{ii}^{1/2}}\right)\theta_1 - \left(\frac{\lambda_{2j}}{\lambda_{jj}^{1/2}} - \frac{\lambda_{2i}}{\lambda_{ii}^{1/2}}\right)\theta_2\right]$$

$$= \left[(\rho_{1j}-\rho_{1i})\,\lambda_{11}^{1/2}\theta_1 - (\rho_{2j}-\rho_{2i})\,\lambda_{22}^{1/2}\theta_2\right]/[2(1-\rho_{ij})]^{1/2},$$

$$\text{Var}\,(t_{ij}^*) = \sigma^2\,\underset{\sim}{d}'\,\underset{\sim}{d} = \sigma^2,$$

on using equation (5.3.5).

Next, $Q_1 = t_{ij}^{*2} = \underset{\sim}{y}'\,\underset{\sim}{d}\,\underset{\sim}{d}'\,\underset{\sim}{y}$.

Hence $Q_1 \overset{d}{=} \chi^2(1,\delta^*)$, where the non-centrality parameter is given by

$$\sigma^2\delta^* = [E(\underset{\sim}{y})]'\,\underset{\sim}{d}\,\underset{\sim}{d}'\,[E(\underset{\sim}{y})]$$

$$= [E(t_{ij}^*)]^2 = \mu^{*2}.$$

Further $Q_2 = S^2 - t_{ij}^{*2}$

$$= \underset{\sim}{y}'\,\underset{\sim}{\Lambda}\,\underset{\sim}{y} - \underset{\sim}{y}'\,\underset{\sim}{d}\,\underset{\sim}{d}'\,\underset{\sim}{y}$$

$$= \underset{\sim}{y}'[\underset{\sim}{\Lambda} - \underset{\sim}{d}\,\underset{\sim}{d}']\,\underset{\sim}{y}.$$

Thus Q_2 is also a quadratic form in $\underset{\sim}{y}$.

The matrix of the quadratic form is $\underset{\sim}{\Lambda} - \underset{\sim}{d}\,\underset{\sim}{d}'$, which satisfies

$$(\underset{\sim}{\Lambda} - \underset{\sim}{d}\,\underset{\sim}{d}')^2 = \underset{\sim}{\Lambda}^2 - \underset{\sim}{\Lambda}\,\underset{\sim}{d}\,\underset{\sim}{d}' + \underset{\sim}{d}\,\underset{\sim}{d}'\,\underset{\sim}{d}\,\underset{\sim}{d}' - \underset{\sim}{d}\,\underset{\sim}{d}'\underset{\sim}{\Lambda}$$

$$= \underset{\sim}{\Lambda} - \underset{\sim}{\Lambda}\,\underset{\sim}{d}\,\underset{\sim}{d}' + \underset{\sim}{d}\,\underset{\sim}{d}' - \underset{\sim}{d}\,\underset{\sim}{d}'\underset{\sim}{\Lambda}$$

on using equation (5.3.5). Further

$$\underset{\sim}{\Lambda}_2 \underset{\sim}{d}_2 = \begin{bmatrix} \underset{\sim}{\lambda}_2'(1) \\ \underset{\sim}{\lambda}_2'(2) \\ \vdots \\ \underset{\sim}{\lambda}_2'(n) \end{bmatrix} \frac{1}{[2(1-\rho_{ij})]^{1/2}} \left[\frac{\underset{\sim}{\lambda}_2(i)}{\lambda_{ii}^{1/2}} - \frac{\underset{\sim}{\lambda}_2(j)}{\lambda_{jj}^{1/2}}\right]$$

$$= \frac{1}{[2(1-\rho_{ij})]^{1/2}} \begin{bmatrix} \lambda_{1i}/\lambda_{ii}^{1/2} - \lambda_{1j}/\lambda_{jj}^{1/2} \\ \lambda_{2i}/\lambda_{ii}^{1/2} - \lambda_{2j}/\lambda_{jj}^{1/2} \\ \vdots \\ \lambda_{ni}/\lambda_{ii}^{1/2} - \lambda_{nj}/\lambda_{jj}^{1/2} \end{bmatrix}$$

$$= \frac{1}{[2(1-\rho_{ij})]^{1/2}} \left[\frac{\underset{\sim}{\lambda}_2(i)}{\lambda_{ii}^{1/2}} - \frac{\underset{\sim}{\lambda}_2(j)}{\lambda_{jj}^{1/2}}\right]$$

$$= \underset{\sim}{d}_2.$$

Consequently $(\underset{\sim}{\Lambda}_2 - \underset{\sim}{d}_2 \underset{\sim}{d}_2')^2 = \underset{\sim}{\Lambda}_2 - \underset{\sim}{d}_2 \underset{\sim}{d}_2'$, that is $\underset{\sim}{\Lambda}_2 - \underset{\sim}{d}_2 \underset{\sim}{d}_2'$ is an idempotent matrix of rank given by

$$\begin{aligned} \text{rank}\,(\underset{\sim}{\Lambda}_2 - \underset{\sim}{d}_2 \underset{\sim}{d}_2') &= \text{tr}\,(\underset{\sim}{\Lambda}_2 - \underset{\sim}{d}_2 \underset{\sim}{d}_2') \\ &= \text{tr}\,\underset{\sim}{\Lambda}_2 - \text{tr}\,(\underset{\sim}{d}_2 \underset{\sim}{d}_2') \\ &= \text{rank}\,(\underset{\sim}{\Lambda}_2) - \text{tr}\,(\underset{\sim}{d}_2' \underset{\sim}{d}_2) \\ &= n-k-\underset{\sim}{d}_2'\underset{\sim}{d}_2 = n-k-1. \end{aligned}$$

Thus $Q_2 \overset{d}{=} \sigma^2 \chi^2(n-k-1, \eta^*)$,

where

$$\begin{aligned} \sigma^2\eta^* &= [\,E(\underset{\sim}{y}_2)]'\,[\underset{\sim}{\Lambda}_2 - \underset{\sim}{d}_2 \underset{\sim}{d}_2']\,[E(\underset{\sim}{y}_2)] \\ &= E(\underset{\sim}{y}_2')\ \underset{\sim}{\Lambda}_2\ E(\underset{\sim}{y}_2) - E(\underset{\sim}{y}_2')\ \underset{\sim}{d}_2 \underset{\sim}{d}_2'\ E(\underset{\sim}{y}_2) \\ &= \sigma^2\ \lambda^{**} - \sigma^2\ \delta^*, \end{aligned}$$

where λ^{**} is given in equation (5.3.1) with i = 1 and j = 2, and $\sigma^2 \delta^* = \mu^{*2}$. Hence

$$\sigma^2 \eta^* = \lambda_{11}\theta_1^2 + \lambda_{22}\theta_2^2 - 2\lambda_{12}\theta_1\theta_2$$

$$- [1/\{2(1-\rho_{ij})\}] [(\rho_{1j}-\rho_{1i})^2 \lambda_{11}\theta_1^2 + (\rho_{2j}-\rho_{2i})^2 \lambda_{22}\theta_2^2$$

$$- 2(\rho_{1j}-\rho_{1i})(\rho_{2j}-\rho_{2i})(\lambda_{11}\lambda_{22})^{1/2}\theta_1\theta_2]$$

$$= [1/\{2(1-\rho_{ij})\}] [\{2(1-\rho_{ij}) - (\rho_{1j}-\rho_{1i})^2\}\lambda_{11}\theta_1^2$$

$$+ \{2(1-\rho_{ij}) - (\rho_{2j}-\rho_{2i})^2\}\lambda_{22}\theta_2^2$$

$$- 2\{2\lambda_{12}(1-\rho_{ij}) - (\rho_{1j}-\rho_{1i})(\rho_{2j}-\rho_{2i})(\lambda_{11}\lambda_{22})^{1/2}\}\theta_1\theta_2] .$$

Finally, $t_{ij}^* = \underset{\sim}{d}' \underset{\sim}{y}$ and $Q_2 = \underset{\sim}{y}'(\underset{\sim}{\Lambda} - \underset{\sim}{d}\,\underset{\sim}{d}')\,\underset{\sim}{y}$ are independent, since

$$(\underset{\sim}{\Lambda} - \underset{\sim}{d}\,\underset{\sim}{d}')\,\underset{\sim}{d} = \underset{\sim}{\Lambda}\,\underset{\sim}{d} - \underset{\sim}{d}\,\underset{\sim}{d}'\,\underset{\sim}{d} = \underset{\sim}{d} - \underset{\sim}{d} = \underset{\sim}{0}.$$

Consequently $Q_1 = t_{ij}^{*2}$ and Q_2 are also independent.

This completes the proof of the lemma.

Now using these values of $\mu^{**} = \mu^*/\sigma$, η^* and δ^*, and Theorem 5.2.1, we get the exact non-null distribution of v_{ij} under H_{12}^* as

$$(5.3.6)\quad f(v_{ij}) = \sum_{j_1=0}^{\infty} K_{j_1} \sum_{i_1=0}^{\infty} K_{i_1}^* \frac{(v_{ij})^{i_1}(1-v_{ij}^2)^{j_1+a/2-1}}{B[(i_1+1)/2, j_1+a/2]},$$

$$-1 \le v_{ij} \le 1,$$

where

$$(5.3.7) \qquad K_{j_1} = e^{-\eta^*/2} (\eta^*/2)^{j_1}/j_1! \quad , \; j_1 = 0,1,2,\ldots,$$

$$(5.3.8) \qquad K^*_{i_1} = e^{-\delta^*/2} (\delta^*/2)^{i_1/2}/G(i_1/2+1), \; i_1 = 0,1,2,\ldots,$$

$$\delta^* = \mu^{**2} = \mu^{*2}/\sigma^2$$

and $a = n-k-1$.

For studying the performance of the statistic V, we are interested in obtaining (for $v_\alpha > 0$)

$$\Pr(V > v_\alpha | H_{12}^*) = \Pr\left(\max_{1 \le i < j \le n} |v_{ij}| > v_\alpha | H_{12}^* \right)$$

$$\le \sum_{1 \le i < j \le n} \Pr(|v_{ij}| > v_\alpha | H_{12}^*)$$

$$= \sum_{1 \le i < j \le n} \Pr(v_{ij}^2 > v_\alpha^2 | H_{12}^*).$$

Hence it is sufficient to study the distribution of v_{ij}^2 for obtaining a measure of performance of V. The following theorem gives an approximate distribution of v_{ij}^2. This is obtained by using Patnaik's (1949) approximation for non-central χ^2 distribution.

Theorem 5.3.1. Let Z_1 and Z_2 be independently distributed as $\chi^2(a_1, \lambda_1)$ and $\chi^2(a_2, \lambda_2)$ respectively. Then the pdf of

$$(5.3.9) \qquad Z = Z_1/(Z_1 + Z_2)$$

can be approximated by

$$(5.3.10)\quad f(z) = \frac{c_1^{f_2/2}\, c_2^{f_1/2}\, z^{f_1/2-1}\,(1-z)^{f_2/2-1}}{B(f_1/2, f_2/2)\,[c_1+(c_2-c_1)z]^{(f_1+f_2)/2}},\quad 0 \le z \le 1,$$

where for $i = 1,2$,

$$(5.3.11)\quad c_i = (a_i + 2\lambda_i)/(a_i + \lambda_i) \text{ and}$$

$$(5.3.12)\quad f_i = (a_i + \lambda_i)^2/(a_i + 2\lambda_i).$$

Proof : Since $Z_i \overset{d}{=} \chi^2(a_i, \lambda_i)$, $(i = 1,2)$, hence each Z_i can be approximated by Patnaik's approximation, for example, see Johnson and Kotz (1970, p.198), as $Z_i = c_i\, Y_i$, where Y_i is distributed as a central χ^2 with f_i degrees of freedom, and c_i and f_i are defined in equations (5.3.11) and (5.3.12) respectively.

The pdf of Y_i is given by

$$f(y_i) = e^{-y_i/2}\, y_i^{f_i/2-1}/[2^{f_i/2}\, G(f_i/2)]\ ,\ 0 < y_i \le \infty,\ i = 1,2\ .$$

Hence the pdf of Z_i is

$$f(z_i) = e^{-z_i/(2c_i)}\, z_i^{(f_i/2-1)}/[(2c_i)^{f_i/2}\, G(f_i/2)]\ ,$$

$$0 < z_i \le \infty,\ i = 1,2.$$

Since Z_1 and Z_2 are independent, hence the joint pdf of Z_1 and Z_2 is given by

$$f(z_1, z_2) = \frac{e^{-\{z_1/(2c_1)+z_2/(2c_2)\}}\, z_1^{f_1/2-1}\, z_2^{f_2/2-1}}{2^{(f_1+f_2)/2}\, c_1^{f_1/2}\, c_2^{f_2/2}\, G(f_1/2)\, G(f_2/2)}\ .$$

Now making a transformation

$$z = z_1/(z_1+z_2) \text{ and } x = z_1+z_2,$$

$$0 \leq z \leq 1,\ 0 \leq x \leq \infty,$$

the inverse transformation is given by

$$z_1 = zx,\ z_2 = x(1-z).$$

The jacobian of transformation $\left|\frac{\partial(z_1,z_2)}{\partial(z,x)}\right| = x.$

Hence the joint pdf of Z and X is given by

$$f(z,x) = \frac{z^{f_1/2-1}(1-z)^{f_2/2-1}\, e^{-\{z/(2c_1)+(1-z)/(2c_2)\}x}\, x^{(f_1+f_2)/2-1}}{2^{(f_1+f_2)/2}\, c_1^{f_1/2}\, c_2^{f_2/2}\, G(f_1/2)\, G(f_2/2)}.$$

Integrating x from 0 to ∞,

$$f(z) = \frac{G[(f_1+f_2)/2]\, z^{f_1/2-1}(1-z)^{f_2/2-1}}{2^{(f_1+f_2)/2}\, G(f_1/2)\, G(f_2/2)\, c_1^{f_1/2}\, c_2^{f_2/2}\, [z/(2c_1)+(1-z)/(2c_2)]^{(f_1+f_2)/2}}$$

$$= \frac{c_1^{f_2/2}\, c_2^{f_1/2}\, z^{f_1/2-1}(1-z)^{f_2/2-1}}{B(f_1/2,f_2/2)\, [c_1+(c_2-c_1)z]^{(f_1+f_2)/2}},\ 0 \leq z \leq 1.$$

This completes the proof of the theorem.

Note that f(z) satisfies the properties of a density function, that is (i) $f(z) \geq 0$ and (ii) $\int_0^1 f(z) = 1$. This can be verified as follows.

Since $c_1, c_2 \geq 0$ and $0 < z < 1$, hence $f(z) \geq 0$.

Now, to check the second property, let

$$z_1 = c_2 z/[c_1 + (c_2 - c_1)z] .$$

Then $1-z_1 = c_1(1-z)/[c_1+(c_2-c_1)z]$

and $\frac{dz_1}{dz} = c_1c_2/[c_1+(c_2-c_1)z]^2.$

Hence

$$\int_0^1 f(z)dz$$

$$= \frac{1}{B(f_1/2,f_2/2)} \int_0^1 \frac{(c_2z)^{f_1/2-1}[c_1(1-z)]^{f_2/2-1} c_1c_2\, dz}{[c_1+(c_2-c_1)z]^{f_1/2}\,[c_1+(c_2-c_1)z]^{f_2/2}}$$

$$= \frac{1}{B(f_1/2,f_2/2)} \int_0^1 z_1^{f_1/2-1}(1-z_1)^{f_2/2-1}\, dz_1$$

$$= 1,$$

that is, $f(z)$ is indeed a density function.

The results of Theorem 5.3.1 can now be applied for obtaining an approximate pdf of v_{ij}^2.

Similar to the expression of u_{ij} in equation (5.1.1), the statistic v_{ij} of equation (2.1.1) also reduces to

$$(5.3.13)\quad v_{ij} = (e_i/\lambda_{ii}^{1/2} - e_j/\lambda_{jj}^{1/2})/[S\{2(1-\rho_{ij})\}^{1/2}] = t_{ij}^*/S,$$

on using equation (5.3.2). Hence

$$v_{ij}^2 = t_{ij}^{*2}/S^2 = t_{ij}^{*2}/(t_{ij}^{*2} + Q_2)$$

$$= \frac{Q_1/\sigma^2}{(Q_1/\sigma^2 + Q_2/\sigma^2)} .$$

By Lemma 5.3.1, we see that under H_{12}^*, $Q_1 \overset{d}{=} \sigma^2 \chi^2(1,\delta^*)$ and $Q_2 \overset{d}{=} \sigma^2\chi^2(n-k-1,\eta^*)$, where $\delta^* = \mu^{*2}/\sigma^2$, μ^* and η^* are as defined by equations (5.3.3) and (5.3.4) respectively.

Further Q_1 and Q_2 are independent. The conditions of the Theorem 5.3.1 are satisfied if we take $a_1 = 1$, $a_2 = n-k-1$, $\lambda_1 = \delta^*$ and $\lambda_2 = \eta^*$. Hence we have the following theorem.

Theorem 5.3.2. An approximate non-null pdf of $v^* = v_{ij}^2$ under H_{12}^* is given by

(5.3.14) $f(v^*)$

$$= \frac{c_1^{f_2/2}\, c_2^{f_1/2}\, (v^*)^{f_1/2-1}\, (1-v^*)^{f_2/2-1}}{B(f_1/2, f_2/2)\, [c_1+(c_2-c_1)v^*]^{(f_1+f_2)/2}}, \quad 0 \le v^* \le 1,$$

where $c_i = (a_i + 2\lambda_i)/(a_i + \lambda_i)$, $f_i = (a_i+\lambda_i)^2/(a_i+2\lambda_i)$; $i=1,2$; $a_1 = 1$, $a_2 = n-k-1$, $\lambda_1 = \delta^*$ and $\lambda_2 = \eta^*$.

Note that $\theta_1 = \theta_2 = 0$ implies that $\delta^* = \eta^* = 0$. Consequently $c_1 = c_2 = 1$ and $f_1 = a_1 = 1$, $f_2 = a_2 = n-k-1 = p-1$. This reduces equation (5.3.14) to

$$f(v^*) = \frac{1}{B\,[1/2,(p-1)/2]}\, v^{*-1/2}\, (1-v^*)^{(p-3)/2}, \quad 0 \le v^* \le 1;$$

which agrees with the exact null distribution of $v^* = v_{ij}^2$ obtained from equation (2.2.11). Thus in the null case, the approximate distribution given here coincides with the exact distribution.

Corollary 5.3.1. The probability

$$p^{*i,j} = \Pr(|v_{ij}| > v_\alpha | H_{12}^*)$$

$$\cong 1-I_z(f_1/2,\ f_2/2), \tag{5.3.15}$$

where

$$z = c_2 v_\alpha^2/[c_1+(c_2-c_1)\ v_\alpha^2]$$

$$c_i = (a_i+2\lambda_i)/(a_i+\lambda_i)$$

$$f_i = (a_i+\lambda_i)^2/(a_i+2\lambda_i),\ i = 1,2;$$

$$a_1 = 1,\ a_2 = n-k-1,\ \lambda_1 = \delta^*\ \text{and}\ \lambda_2 = \eta^*.$$

Proof : $p^{*i,j} = \Pr(|v_{ij}| > v_\alpha | H_{12}^*)$

$$= \Pr(v_{ij}^2 > v_\alpha^2 | H_{12}^*)$$

$$= 1-\Pr(v_{ij}^2 \leq v_\alpha^2 | H_{12}^*)$$

$$\cong 1-\int_0^{v_\alpha^2} \frac{c_1^{f_1/2}\ c_2^{f_2/2}\ v^{*f_1/2-1}\ (1-v^*)^{f_2/2-1}}{B(f_1/2,f_2/2)[(c_1+(c_2-c_1)v^*]^{(f_1+f_2)/2}}\, dv^*.$$

Let $z_1 = c_2\ v^*/[\ c_1+(c_2-c_1)v^*]$, then

$$1 - z_1 = c_1(1-v^*)/[c_1+(c_2-c_1)\ v^*]$$

and $\dfrac{dz_1}{dv^*} = c_1c_2/[c_1+(c_2-c_1)v^*]^2.$

Hence

$$p^{*i,j} \cong 1-\int_0^{c_2 v_\alpha^2/[c_1+(c_2-c_1)v_\alpha^2]} \frac{z_1^{f_1/2-1}\ (1-z_1)^{f_2/2-1}}{B(f_1/2,f_2/2)}\, dz_1$$

$$= 1-I_z(f_1/2,f_2/2),\ \text{where}\ z = c_2 v_\alpha^2/[c_1+(c_2-c_1)v_\alpha^2].$$

Note that $\theta_1 = \theta_2 = 0$ implies $f_1 = 1$, $f_2 = p-1$ and $z = v_\alpha^2$. Therefore, under H_o,

$$p^{*i,j} = Pr(|v_{ij}| > v_\alpha | H_o) = 1 - I_{v_\alpha^2}[1/2, (p-1)/2]$$

$$= I_{1-v_\alpha^2}[(p-1)/2, 1/2]$$

$$= 2\alpha/[n(n-1)] = \alpha/\binom{n}{2},$$

from equation (2.3.2), the value in the null case.

The exact method which was discussed for u_{ij}'s in Section 5.2 can also be applied for v_{ij}'s. The distribution of v_{ij} under H_{12}^* is given at equation (5.3.6). Further

$$p^{*i,j} = Pr(|v_{ij}| > v_\alpha | H_{12}^*)$$

$$= Pr(v_{ij}^2 > v_\alpha^2 | H_{12}^*).$$

Equation (5.2.13) now gives

$$p^{*i,j} = \sum_{j=0}^{\infty} K_j \sum_{i=0}^{\infty} K_{2i}^* \int_{v_\alpha^2}^{1} \frac{v^{(i-1/2)} (1-v)^{j+a/2-1} \, dv}{B(i+1/2, j+a/2)}$$

$$(5.3.16) \quad = \sum_{j=0}^{\infty} K_j \sum_{i=0}^{\infty} K_{2i}^* I_{1-v_\alpha^2}[j+a/2, i+1/2],$$

where $a = n-k-1$, K_j and K_i^* are as defined in equations (5.3.7) and (5.3.8) respectively.

Table 5.3.1 gives the exact and approximate values of $p^{*1,2}$ for a random sample of size $n = 10$ from $N(\mu, \sigma^2)$ distribution for different combinations of θ_1 and θ_2 and for $\alpha = 0.05$, where $\sigma = 1$

is taken without any loss of generality. The relative error = $1-p^{*1,2}$ approx./$p^{*1,2}$ exact, is also tabulated.

This table shows that the approximate expression for $p^{*1,2}$ given in Corollary 5.3.1 is satisfactory. Similar results hold for other $p^{*i,j}$'s also. Further, the approximate method is economical both costwise and timewise. Consequently it is used for evaluation of these probabilities in later sections.

Approximation for $Pr(u_{ij} > u_\alpha | H_{12})$: We now obtain an approximate expression for $Pr(u_{ij} > u_\alpha | H_{12})$. This is needed, since due to rounding errors the exact method is difficult to deal with, when θ_1 and θ_2 are large.

In equation (5.3.15), if we take $z = c_2 u_\alpha^2/[c_1+(c_2-c_1)u_\alpha^2]$, $c_i = (a_i+2\lambda_i)/(a_i+\lambda_i)$, $f_i = (a_i+\lambda_i)^2/(a_i+2\lambda_i)$ for $i = 1,2$, $a_1 = 1$, $a_2 = n-k-1$, $\lambda_1 = \delta = \mu^2/\sigma^2$ and $\lambda_2 = \eta$, where μ and η are defined in equations(5.2.4) and (5.2.5), then we get an approximation of $Pr(u_{ij}^2 > u_\alpha^2 | H_{12})$, that is $Pr(|u_{ij}| > u_\alpha | H_{12})$. As shown in Section 5.2, the non-null distribution of u_{ij} is essentially confined to interval $[0,1]$ for large values of μ/σ and hence of θ_1 and θ_2. Consequently,

$$p^{i,j} = Pr(u_{ij} > u_\alpha | H_{12}) \tilde{=} Pr(u_{ij}^2 > u_\alpha^2 | H_{12}).$$

Exact expression for $Pr(u_{ij} > u_\alpha | H_{12})$ is given in Corollary 5.2.2. Table 5.3.2 compares the exact and approximate expression for $p^{1,2}$ for a random sample of size n = 10 from $N(\mu,\sigma^2)$ distribution for different combinations of θ_1 and θ_2

and for $\alpha = 0.05$, where we again take $\sigma = 1$. The relative error is also tabulated. It can be seen that the approximation is not good for small values of θ_1 and θ_2. In particular it is quite bad for $\theta_1 = \theta_2 = 0$. However, it is fairly accurate for large values of θ_1 and θ_2. Further, the relative error keeps on decreasing, as θ_1, θ_2 increase. Similar conclusions hold for other $P^{i,j}$'s. Due to excessive cost and time consumed for evaluating exact value for $P^{i,j}$, the approximate method is used for large values of θ_1 and θ_2 in our performance studies of the statistic U.

5.4. Measures of performance

We now study the performance of test statistic U and V under suitable alternative hypotheses. Theoretically, the best measure of performance is the power function of the test. However, this is extremely difficult to evaluate. We therefore concentrate on some simple and easy to calculate measures of performance. Our measures of performance are analogous to the measures proposed and studied by David and Paulson (1965), McMillan (1971) and Joshi (1972). We also assume that a priori every pair of observations has an equal chance of being an outlying pair.

5.4.1. Measures of performance of the statistic U

For this case, the alternative hypothesis is as specified in Section 5.2. For $1 \leq i < j \leq n$, let

(5.4.1) $\quad P_{ij} = Pr(u_{ij} > u_\alpha | H_{ij})$, and

(5.4.2) $$Q_{ij} = \Pr(U > u_\alpha | H_{ij}).$$

Since $U = \underset{1\le i<j\le n}{\text{Max}} u_{ij}$, hence we obviously have $P_{ij} \le Q_{ij}$. Both P_{ij} and Q_{ij} are reasonable measures of performance. However, one may go a step further, and consider

(5.4.3) $$P_a = \underset{1\le i<j\le n}{\text{Min}} P_{ij}, \text{ and}$$

(5.4.4) $$Q_a = \underset{1\le i<j\le n}{\text{Min}} Q_{ij}.$$

If P_{ij} and Q_{ij} do not depend on i and j, then $P_a = P_{12}$ is the probability that u_{12} is significantly large when the alternative hypothesis is true. Similarly $Q_a = Q_{12}$ is the power function.

Numerical values for P_{ij} can be calculated exactly from equation (5.2.15). They can also be approximated by the technique discussed at the end of Section 5.3. However, exact evaluation of Q_{ij} is too complicated. But the first Bonferroni inequality can be used to get an upper bound for Q_{ij}. Thus

$$\begin{aligned} Q_{ij} &= \Pr(U > u_\alpha | H_{ij}) \\ &= \Pr\Big(\underset{1\le i_1<j_1\le n}{\text{Max}} u_{i_1 j_1} > u_\alpha | H_{ij} \Big) \\ &\le \sum_{1\le i_1<j_1\le n} \Pr(u_{i_1 j_1} > u_\alpha | H_{ij}), \end{aligned} \tag{5.4.5}$$

where each term of the sum can be obtained from the non-null distribution obtained in Theorem 5.2.2. Obviously an upper limit for Q_{ij} is 1. Hence

$$Q_{ij} \le \text{Min} \Big[1, \sum_{1\le i_1<j_1\le n} \Pr(u_{i_1 j_1} > u_\alpha | H_{ij})\Big].$$

5.4.2. Measures of performance of the statistic V

Measure of performance for V can be defined in a manner similar to that of U. Now the alternative hypothesis is as described in Section 5.3. For $i \neq j = 1,2,\ldots,n$, let

$$P^*_{ij} = \Pr(|v_{ij}| > v_\alpha | H^*_{ij}), \tag{5.4.6}$$

$$Q^*_{ij} = \Pr(V > v_\alpha | H^*_{ij}). \tag{5.4.7}$$

We also consider the measures

$$P^*_a = \min_{i \neq j} P^*_{ij}, \text{ and} \tag{5.4.8}$$

$$Q^*_a = \min_{i \neq j} Q^*_{ij}. \tag{5.4.9}$$

Again if P^*_{ij} and Q^*_{ij} do not depend on i and j, then P^*_a and Q^*_a have similar interpretations as for P_a and Q_a defined for statistic U. As before, we have

$$P^*_{ij} \leq Q^*_{ij},$$

and
$$Q^*_{ij} = \Pr\left(\max_{1 \leq i_1 < j_1 \leq n} |v_{i_1 j_1}| > v_\alpha | H^*_{ij} \right)$$

$$\leq \sum_{1 \leq i_1 < j_1 \leq n} \Pr(|v_{i_1 j_1}| > v_\alpha | H^*_{ij}). \tag{5.4.10}$$

The non-null probabilities appearing in equations (5.4.6) and (5.4.10) can be obtained exactly by using equation (5.3.16).

5.4.3. Application to a random sample from $N(\mu,\sigma^2)$ distribution

We now apply our measures of performance to the case of a random sample of size n from $N(\mu,\sigma^2)$ distribution. Since all measures under H_{ij} depend on the ratios θ_i/σ and θ_j/σ, hence

we take $\sigma = 1$ without loss of generality.

For the statistic U, now P_{ij} and Q_{ij} do not depend on i and j. Consequently $Q_{12} = Q_a$ and it is possible to evaluate lower and upper bounds for the power function. A lower bound is of course P_{12}, while an upper bound is given by

$$Q_{12} \leq \text{Min}\ (\bar{Q}_{12}, 1), \tag{5.4.11}$$

where from equation (5.4.5), we have

$$\bar{Q}_{12} = \Pr(u_{12} > u_\alpha | H_{12}) + (n-2)\ \Pr(u_{13} > u_\alpha | H_{12})$$

$$+ (n-2)\ \Pr(u_{23} > u_\alpha | H_{12}) + \binom{n-2}{2}\ \Pr(u_{34} > u_\alpha | H_{12}).$$

For notational convenience, denote

$\Pr(u_{ij} > u_\alpha | H_{12})$ by $P^{i,j}$, where the dependence on H_{12} is suppressed. In this notation $P^{1,2} = P_{12}$ of equation (5.4.1). Consequently,

$$\bar{Q}_{12} = P^{1,2} + (n-2)\ (P^{1,3} + P^{2,3}) + \binom{n-2}{2}\ P^{3,4}. \tag{5.4.12}$$

The bound (5.4.11) is useful when $\bar{Q}_{12} < 1$, which is the case for small values of n and α.

As shown in Section 3.1, the test for two outliers based on U is equivalent to the Murphy's test. For Murphy's test, the measure P_{12} has been studied by McMillan (1971), Moran and McMillan (1973), while an approximate method for evaluating power Q_{12} is given by Hawkins (1978). Hawkins has tabulated approximate power for n = 10 and $\alpha = 0.05$. In Table 5.4.1, we tabulate $P^{1,2}$, $P^{1,3}$, $P^{2,3}$, $P^{3,4}$ and min $(\bar{Q}_{12}, 1)$ for n = 10 and

$\alpha = 0.05$, evaluated by the exact formula for $\theta_1, \theta_2 = 0,1,\ldots,5$. For $\theta_1, \theta_2 \leq 7$, lower and upper limits for power function Q_{12} are given in Table 5.4.2 by approximate method. As pointed out in Section 5.3 the approximation is not good for small values of θ_1 and θ_2. For the sake of comparison, we have also evaluated these probabilities by using Monte Carlo techniques. The method followed is as follows :

A sample of size 10 from a standard normal population is generated. Then the values θ_1 and θ_2 are added to the first and the second observation of the sample. The values u_{ij} are calculated and compared with the nominal percentile point u_α obtained from Table 2.3.1. This procedure is repeated N = 10,000 times and the total number of times each of these u_{ij}'s exceeds u_α value are counted. The number of times the maximum of these u_{ij}'s exceeds u_α value is also obtained. Since

$$P^{1,2} = \Pr(u_{12} > u_\alpha | H_{12}),$$

hence $P^{1,2}$ is calculated by

$$P^{1,2} = \frac{\text{Number of times } (u_{12} > u_\alpha) \text{ in N repetitions}}{N}.$$

Similarly

$$Q^{1,2} = \frac{\text{Number of times } (U > u_\alpha) \text{ in N repetitions}}{N}.$$

However, $P^{1,3} = P^{1,4} = \ldots = P^{1,10}$, hence we evaluate $P^{1,3}$ by counting the total number of times $(u_{1j} > u_\alpha)$ for $j = 3,4,\ldots,10$ in N repetitions and dividing this number by 8N. Similarly, using the facts that $P^{2,j}$ for $j = 3,4,\ldots,10$ and $P^{i,j}$ for $3 \leq i < j \leq 10$

are all equal, we can evaluate $P^{2,3}$ and $P^{3,4}$. This is repeated for different values of θ_1 and θ_2. These values are tabulated in Table 5.4.3.

A comparison of Tables 5.4.1 and 5.4.3 reveals that the upper bound for power given in equations (5.4.11) and (5.4.12) is quite close to the simulated value for $\theta_1+\theta_2 \geq 4$. In fact, $P^{1,2}$ itself is quite close to the power value for $\theta_1,\theta_2 \geq 3$. Consequently, one can evaluate the power function approximately by finding $P^{1,2}$ alone or by finding $\bar{Q}^{1,2}$, for which some extra computations are required. We here point out that the approximate power values of Murphy's test given by Hawkins (1978) for these values of n and α are much less than our values. In fact, our lower bound itself is considerably larger than the power values tabulated by him. For example, our values for $P^{1,2}$ and his values within brackets for $\theta_1 = \theta_2 = \theta$ are 0.704(0.594), 0.911(0.844) for θ = 4 and 5 respectively. This could be due to the fact that he assumes a conditional probability term of his expression as equal to 1, which may not be the case.

Similarly for $\theta_1 = 3$, $\theta_2 = 5$ we have $P^{1,2} = 0.574$, while he tabulates 0.481 as power of the test.

A method for finding $P^{1,2}$ for Murphy's test is also given by McMillan (1971) using non-central t-distribution. However, he has not provided any tables, but has provided a curve for n = 11 and α = 0.02625. Our values of $P^{1,2}$ for this combination of n and α compare favourably with his values.

As pointed out by McMillan (1971) and Hawkins (1978), Murphy's test performs well if θ_1 and θ_2 are approximately equal. However, its power deteriorates if θ_1 is markedly different from θ_2. Our calculations also justify their statement. In fact, if $\theta_1 = 0$, that is, if there is only one outlier then the power of the Murphy's test (for n = 10 and $\alpha = 0.05$) first increases, attains a maximum value around $\theta_2 = 4$, and then decreases to zero.

For the statistic V, which is equivalent to Studentized range test statistic, P^*_{ij} and Q^*_{ij} do not depend on i and j. Consequently $Q^*_{12} = Q^*_a$ and it is possible to evaluate lower and upper bounds for the power function. Now P^*_{12} is a lower bound, while an upper bound is given by

$$(5.4.13) \qquad Q^*_{12} \leq \text{Min } (\bar{Q}^*_{12}, 1),$$

where from equation (5.4.10) we have

$$\bar{Q}^*_{12} = \Pr(|v_{12}| > v_\alpha | H^*_{12}) + (n-2)\Pr(|v_{13}| > v_\alpha | H^*_{12})$$
$$+(n-2)\Pr(|v_{23}| > v_\alpha | H^*_{12}) + \binom{n-2}{2}\Pr(|v_{34}| > v_\alpha | H^*_{12}).$$

For notational convenience, we denote $\Pr(|v_{ij}| > v_\alpha | H^*_{12})$ by $p^{*i,j}$. In this notations $p^{*1,2} = P^*_{12}$ of equation (5.4.6). Consequently,

$$(5.4.14) \qquad \bar{Q}^*_{12} = p^{*1,2} + (n-2)\,(p^{*1,3} + p^{*2,3}) + \binom{n-2}{2}\, p^{*3,4}.$$

Similar to the case of U, we tabulate $p^{*1,2}$, $p^{*1,3}$, $p^{*2,3}$, $p^{*3,4}$ and $\min(\bar{Q}^*_{12}, 1)$ for n = 10 and $\alpha = 0.05$ evaluated by the

exact formula for $\theta_1,\theta_2 = 0,1,\ldots,5$ in Table 5.4.4. Bounds for power function calculated by approximate method are given in Table 5.4.5 for $\theta_1,\theta_2 \leq 7$. Although for individual terms $P^{*i,j}$, the approximation is quite satisfactory, yet for small values of θ_1,θ_2 it is not so good for $\bar{Q}^*_{12}$, since $P^{*1,3}$, $P^{*2,3}$ and $P^{*3,4}$ are multiplied by $(n-2),(n-2)$ and $\binom{n-2}{2}$ respectively. For larger values of θ_1 and θ_2, $P^{*1,2}$ is the dominating term in equation (5.4.14) and the difference between exact and approximate values of $\bar{Q}^{*1,2}$ is small. For example, $\theta_1 = \theta_2 = 4$, exact and approximate values for $\bar{Q}^{*1,2}$ are 0.7117 and 0.7113 respectively. Even for unequal values of θ_1 and θ_2, the difference is not much; for example, for $\theta_1 = 2$, $\theta_2 = 5$, we have the exact value equal to 0.3524, while the approximate value is 0.3425. As stated in Section 5.3, the approximate method is much easier and hence one can evaluate the power by this method for larger values of θ_1 and θ_2. However, for small values of θ_1,θ_2, one may still have to calculate the power by exact method, for obtaining better results. For the sake of comparison, we have evaluated these probabilities by Monte Carlo technique also. The method followed is exactly similar to that of U. The power and other values are tabulated in Table 5.4.6 for $\theta_1,\theta_2=0,1,\ldots,5$. From Table 5.4.4 and Table 5.4.6, it is clear that the upper limit for power function is close to simulated value of power for $\theta_1+\theta_2 \geq 4$.

As pointed out for the case of U, the power of statistic V also deteriorates if θ_1 is markedly different from θ_2. In this

case also if $\theta_1 = 0$, that is, if there is only one outlier, then the power of the Studentized range test (for n = 10 and $\alpha = 0.05$) first increases, attains a maximum value around $\theta_2 = 5$ and then decreases to zero.

5.4.4. Application to a two-way layout

We need the following lemma which is similar to Corollary 5.2.2 for obtaining a measure of performance of the U statistic for a two-way table. Again the measures depend on the ratio θ_i/σ and θ_j/σ. Consequently, we can take $\sigma = 1$ without any loss of generality.

Lemma 5.4.1. For $u_\alpha > 0$,

$$P_{ij} = \Pr(u_{ij} > u_\alpha | H_{ij})$$

$$= \frac{1}{2} \sum_{j_1=0}^{\infty} K_{j_1} \sum_{i_1=0}^{\infty} K^*_{i_1} I_{1-u_\alpha^2}[j_1+a/2, (i_1+1)/2] ,$$

where $a = n-k-1$, K_{j_1} and $K^*_{i_1}$ are as defined in equations (5.2.9) and (5.2.10) with

$$(5.4.15) \qquad \delta = [(1+\rho_{ij})/2] (\lambda_{ii}^{1/2} \theta_i + \lambda_{jj}^{1/2} \theta_j)^2,$$

$$(5.4.16) \qquad \eta = [(1-\rho_{ij})/2] (\lambda_{ii}^{1/2} \theta_i - \lambda_{jj}^{1/2} \theta_j)^2,$$

$\theta_i > 0$ and $\theta_j > 0$ are the deviations of the ith and jth observations from their mean under H_o.

Proof : The proof of this lemma is similar to Theorem 5.2.2. Now, we have from equations (5.1.3) and (5.2.3),

$$\mu = E(t_{ij}|H_{ij}) = \underset{\sim}{c}' E(\underset{\sim}{y}) = \underset{\sim}{c}'(\underset{\sim}{X}\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}_i \theta_i + \underset{\sim}{\varepsilon}_j \theta_j)$$

$$= \underset{\sim}{c}' (\underset{\sim}{\varepsilon}_i \theta_i + \underset{\sim}{\varepsilon}_j \theta_j),$$

where $\underset{\sim}{\varepsilon}_i$ and $\underset{\sim}{\varepsilon}_j$ are the ith and jth column vectors of the identity matrix of order n, and

$$\underset{\sim}{c}' = \frac{1}{[2(1+\rho_{ij})]^{1/2}} \left[\frac{\underset{\sim}{\lambda}'(i)}{\lambda_{ii}^{1/2}} + \frac{\underset{\sim}{\lambda}'(j)}{\lambda_{jj}^{1/2}}\right] . \quad \text{Hence}$$

$$\mu = \frac{1}{[2(1+\rho_{ij})]^{1/2}} \left[\frac{\underset{\sim}{\lambda}'(i)\ \underset{\sim}{\varepsilon}_i\ \theta_i}{\lambda_{ii}^{1/2}} + \frac{\underset{\sim}{\lambda}'(i)\ \underset{\sim}{\varepsilon}_j\ \theta_j}{\lambda_{ii}^{1/2}} + \frac{\underset{\sim}{\lambda}'(j)\ \underset{\sim}{\varepsilon}_i\ \theta_i}{\lambda_{jj}^{1/2}} + \frac{\underset{\sim}{\lambda}'(j)\ \underset{\sim}{\varepsilon}_j\ \theta_j}{\lambda_{jj}^{1/2}}\right]$$

$$= \frac{1}{[2(1+\rho_{ij})]^{1/2}} \left[\left(\lambda_{ii}^{1/2} + \frac{\lambda_{ij}}{\lambda_{jj}^{1/2}}\right) \theta_i + \left(\lambda_{jj}^{1/2} + \frac{\lambda_{ij}}{\lambda_{ii}^{1/2}}\right) \theta_j\right]$$

$$= \frac{1}{[2(1+\rho_{ij})]^{1/2}} \left[(1+\rho_{ij})\ (\lambda_{ii}^{1/2}\ \theta_i + \lambda_{jj}^{1/2}\ \theta_j)\right]$$

$$= \left[(1+\rho_{ij})/2\right]^{1/2}\ (\lambda_{ii}^{1/2}\ \theta_i + \lambda_{jj}^{1/2}\ \theta_j).$$

Since $\delta = \mu^2$, we have the expression for δ as given in (5.4.15). Further, with λ^* as in equation (5.2.1) and $\delta = \mu^2$, we have (for $\sigma = 1$)

$$\eta = \lambda^* - \delta$$

$$= \lambda_{ii}\theta_i^2 + \lambda_{jj}\theta_j^2 + 2\lambda_{ij}\theta_i\theta_j - \frac{(1+\rho_{ij})}{2}(\lambda_{ii}^{1/2}\theta_i + \lambda_{jj}^{1/2}\theta_j)^2$$

$$= \left(1 - \frac{1+\rho_{ij}}{2}\right)(\lambda_{ii}\theta_i^2 + \lambda_{jj}\theta_j^2) + 2\left(\rho_{ij} - \frac{1+\rho_{ij}}{2}\right)(\lambda_{ii}\lambda_{jj})^{1/2}\theta_i\theta_j$$

$$= \left[(1-\rho_{ij})/2\right]\ (\lambda_{ii}^{1/2}\theta_i - \lambda_{jj}^{1/2}\theta_j)^2.$$

On using Theorem 5.2.1 and proceeding as in Corollary 5.2.2, the lemma follows immediately.

Remark 1. If we have, $\lambda_{ii} = \lambda\ (i = 1,2,\ldots,n)$, then the expression of δ and η reduce to

$$\delta = [(1+\rho_{ij})/2]\ \lambda\ (\theta_i+\theta_j)^2, \text{ and}$$

$$\eta = [(1-\rho_{ij})/2]\ \lambda\ (\theta_i-\theta_j)^2.$$

The probability P_{ij} given in Lemma 5.4.1 depends on δ and η apart from 'a' and u_α. For fixed 'a' and u_α, it is not easy to determine theoretically the behaviour of P_{ij} as a function of δ and η, since the weights K_{j_1} and $K^*_{i_1}$, which are like Poisson probability terms, have to be multiplied with incomplete beta integrals. Then the entire series has to be summed over j_1 and i_1 from 0 to ∞. First few terms of this series are given by

$$P_{ij} = \frac{1}{2}\ e^{-(\eta+\delta)/2}\ [b_{o,o} + \frac{\eta}{2}\ b_{o,1} + (\frac{\delta}{2})^{1/2}\ \frac{1}{G(3/2)}\ b_{1,o}+\ldots],$$

where $b_{i_1,j_1} = I_{1-u_\alpha^2}[j_1+a/2,(i_1+1)/2]$, and it is very difficult to determine the behaviour of P_{ij} as a function of δ and η. From limited calculations we have observed that

(i) for fixed η, P_{ij} increases as δ increases, and

(ii) for fixed δ, P_{ij} decreases as η increases.

Now for a two-way table the probability P_{ij} defined in equation (5.4.1) depend on i and j. We therefore consider the measure P_a introduced in equation (5.4.3). The other measure Q_a

defined in equation (5.4.4) is extremely difficult to calculate. For evaluation of P_a we need the following theorem, which is proved by making note of the observation mentioned in Remark 1 above.

Theorem 5.4.1. In a two-way classification with single observation per cell and having r rows and c columns, $P_{ij} = Pr(u_{ij} > u_\alpha | H_{ij})$ is minimum when the two cells in which the two suspected observations are occuring are (1) in the same row or column if $r = c$, (2) in the same column when $r < c$, and (3) in the same row when $r > c$.

Proof : Let the two suspected observations be in the ith and jth cell. From Lemma 5.4.1, we have

$$\delta = [(1+\rho_{ij})/2]\ \lambda\ (\theta_i+\theta_j)^2, \text{ and}$$

$$\eta = [(1-\rho_{ij})/2]\ \lambda\ (\theta_i-\theta_j)^2,$$

where $\lambda = (r-1)(c-1)/rc$ is the common value of λ_{ii} for a two-way table. For a fixed θ_i and θ_j, δ and η would vary according to $(1+\rho_{ij})/2$ and $(1-\rho_{ij})/2$. Thus δ would be minimum and η would be maximum for a minimum value of ρ_{ij}. Since

$$\rho_{ij} = \begin{cases} -1/(c-1) & \text{if the cells i and j are in the same row,} \\ -1/(r-1) & \text{if the cells i and j are in the same column,} \\ 1/[(r-1)(c-1)] & \text{if the cells i and j are neither in the same row nor in the same column,} \end{cases}$$

hence for the minimum value of P_{ij}, the two outlying cells must be in the same row or column according as $r > c$ or $r < c$. For

r = c, the two cells can be in the same row or column. This completes the proof of the theorem.

Now for a measure of performance of the statistic V in case of a two-way table, we require the following lemma, which is analogous to equation (5.3.16).

Lemma 5.4.2. For $v_\alpha > 0$,

$$P^*_{ij} = Pr\,(|v_{ij}| > v_\alpha | H^*_{ij})$$

$$= \sum_{j_1=0}^{\infty} K_{j_1} \sum_{i_1=0}^{\infty} K^*_{2i_1} \, I_{1-v_\alpha^2}[j_1+a/2, i_1+1/2] \,,$$

where $a = n-k-1$, K_{j_1} and $K^*_{i_1}$ are as defined in equations (5.3.7) and (5.3.8) respectively with

$$(5.4.17) \quad \delta^* = [(1-\rho_{ij})/2]\,(\lambda_{ii}^{1/2}\theta_i + \lambda_{jj}^{1/2}\theta_j)^2, \text{ and}$$

$$(5.4.18) \quad \eta^* = [(1+\rho_{ij})/2]\,(\lambda_{ii}^{1/2}\theta_i - \lambda_{jj}^{1/2}\theta_j)^2.$$

Proof : The proof of this lemma is similar to Corollary 5.2.1. Now we have from equations (5.1.3) and (5.3.2)

$$\mu^* = E(t^*_{ij} | H^*_{ij}) = \underset{\sim}{d}' \, E(\underset{\sim}{y})$$

$$= \underset{\sim}{d}' \, (X \underset{\sim}{\beta} - \underset{\sim}{\varepsilon}_i \theta_i + \underset{\sim}{\varepsilon}_j \theta_j)$$

$$= \underset{\sim}{d}' \, (\underset{\sim}{\varepsilon}_j \theta_j - \underset{\sim}{\varepsilon}_i \theta_i),$$

where

$$\underset{\sim}{d}' = \frac{1}{[2(1-\rho_{ij})]^{1/2}} \left[\frac{\underset{\sim}{\lambda}'(i)}{\lambda_{ii}^{1/2}} - \frac{\underset{\sim}{\lambda}'(j)}{\lambda_{jj}^{1/2}}\right]. \quad \text{Hence}$$

$$\mu^* = \frac{1}{[2(1-\rho_{ij})]^{1/2}} \left[\frac{\underset{\sim}{\lambda}'(i)\ \underset{\sim}{\varepsilon}_j\theta_j}{\lambda_{ii}^{1/2}} - \frac{\underset{\sim}{\lambda}'(i)\ \underset{\sim}{\varepsilon}_i\theta_i}{\lambda_{ii}^{1/2}} - \frac{\underset{\sim}{\lambda}'(j)\ \underset{\sim}{\varepsilon}_j\theta_j}{\lambda_{jj}^{1/2}} + \frac{\underset{\sim}{\lambda}'(j)\ \underset{\sim}{\varepsilon}_i\theta_i}{\lambda_{jj}^{1/2}}\right]$$

$$= \frac{1}{[2(1-\rho_{ij})]^{1/2}} \left[\left(\frac{\lambda_{ij}}{\lambda_{jj}^{1/2}} - \lambda_{ii}^{1/2}\right)\theta_i - \left(\lambda_{jj}^{1/2} - \frac{\lambda_{ij}}{\lambda_{ii}^{1/2}}\right)\theta_j\right]$$

$$= -\left[(1-\rho_{ij})/2\right]^{1/2} (\lambda_{ii}^{1/2}\theta_i + \lambda_{jj}^{1/2}\theta_j).$$

Since $\delta^* = \mu^{*2}$, hence

$$\delta^* = \left[(1-\rho_{ij})/2\right] (\lambda_{ii}^{1/2}\theta_i + \lambda_{jj}^{1/2}\theta_j)^2.$$

Further, with λ^{**} as in equation (5.3.1), and $\delta^* = \mu^{*2}$, we have

$$\eta^* = \lambda^{**} - \delta^*$$

$$= \lambda_{ii}\theta_i^2 + \lambda_{jj}\theta_j^2 - 2\lambda_{ij}\theta_i\theta_j - \left[(1-\rho_{ij})/2\right] (\lambda_{ii}^{1/2}\theta_i + \lambda_{jj}^{1/2}\theta_j)^2$$

$$= \left(1 - \frac{1-\rho_{ij}}{2}\right)(\lambda_{ii}\theta_i^2 + \lambda_{jj}\theta_j^2) - 2\left(\rho_{ij} + \frac{1-\rho_{ij}}{2}\right)(\lambda_{ii}\lambda_{jj})^{1/2}\theta_i\theta_j$$

$$= \left[(1+\rho_{ij})/2\right] (\lambda_{ii}^{1/2}\theta_i - \lambda_{jj}^{1/2}\theta_j)^2.$$

On using Theorem 5.2.1 and proceeding as was done for obtaining equation (5.3.16), the lemma follows immediately.

Remark 2. When $\lambda_{ii} = \lambda$, the expression of δ^* and η^* are

$$\delta^* = \left[(1-\rho_{ij})/2\right]\lambda\ (\theta_i+\theta_j)^2$$

$$\eta^* = \left[(1+\rho_{ij})/2\right]\lambda\ (\theta_i-\theta_j)^2.$$

Similar to the case of U, as mentioned in Remark 1, theoretically the behaviour of P^*_{ij} as a function of δ^* and η^* is difficult to determine. Hence again, from limited calculations, we observe that

(i) for fixed η^* , P^*_{ij} increases as δ^* increases, and

(ii) for fixed δ^* , P^*_{ij} decreases as η^* increases.

Now for the measure of performance P^*_a of V given by equation (5.4.8), the alternative hypothesis to be considered is given by the following theorem, which is proved by making note of the observation mentioned in Remark 2 above.

Theorem 5.4.2. In a two-way classification with single observation per cell and having r rows and c columns, the P^*_{ij} value given in Lemma 5.4.2 is minimum when the two cells in which the suspected outliers are lying are neither in the same row nor in the same column.

Proof : Suppose the two outlying observations are lying in the ith and jth cells, then as mentioned in Remark 2, for P^*_{ij} to be minimum, we need δ^* to be minimum and η^* to be maximum. For fixed θ_1 and θ_2, δ^* and η^* would vary according to $(1-\rho_{ij})/2$ and $(1+\rho_{ij})/2$. Thus δ^* would be minimum and η^* would be maximum when ρ_{ij} assumes its maximum value $\frac{1}{(r-1)(c-1)}$. This happens when the two cells in which the suspected observations are lying are neither in the same row nor in the same column. This completes the proof of the theorem.

For the sake of comparison in Section 5.5, we tabulate approximate values of $P_a^* = P^{*11,22}$ for 4x5 and 6x8 tables, for $\alpha = 0.02625$ in Table 5.4.7, for $0 \leq \theta_1 \leq \theta_2 \leq 7$.

Similar to the case of a random sample from normal distribution, here also, the performance of test statistic V is not very good when θ_1 is much smaller compared to θ_2. However, for same values of θ_1, θ_2 the statistic performs better for 6x8 table compared to 4x5 case.

The measure $P_a = P^{11,21}$ (for $r < c$ case) studied for 4x5 and 6x8 tables gives analogous results for the statistic U.

5.5. Comparison with sequential procedure

For comparing the performance of our procedure, we consider the sequential procedure. The procedure for a random sample from $N(\mu,\sigma^2)$ as discussed by McMillan (1971) and McMillan and David (1971) is as follows :

Suppose $Y_1, Y_2, \ldots, Y_n$ are independent, normally distributed observations with common variance σ^2 and in the absence of outliers, (H_o) common mean μ. Let $Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n)}$ be the corresponding order statistics, $\bar{y} = \Sigma y_i/n$ and $S^2 = \Sigma(y_i-\bar{y})^2$. If $Y_{(n)} - \bar{y} > v_\alpha^{(n)} S$, then $Y_{(n)}$ is declared an outlier and the test is repeated using the remaining observations.

If $Y_{(n-1)} - \bar{y}_{(n)} > v_\alpha^{(n-1)} S_{(n)}$, where

$$S^2_{(n)} = \sum_{i=1}^{n-1} (y_{(i)}-\bar{y}_{(n)})^2, \quad \bar{y}_{(n)} = \sum_{i=1}^{n-1} y_{(i)}/(n-1),$$

then $y_{(n-1)}$ is also declared an outlier, etc. Values of $v_\alpha^{(n)}$ such that $\Pr\{y_{(n)}-\bar{y} > v_\alpha^{(n)} S|H_o\} = \alpha$ are given by Quesenberry and David (1961), and related constants are tabulated by Grubbs (1950), and Grubbs and Beck (1972).

To evaluate the performance of this procedure, we suppose that two of the observations, which without loss of generality we take as y_1, y_2, come from a different normal population than the rest of the sample. Specifically, we assume without loss of generality that $y_1 \overset{d}{=} N(\mu+\theta_1, \sigma^2)$ and $y_2 \overset{d}{=} N(\mu+\theta_2, \sigma^2)$, where $\theta_1, \theta_2 > 0$. Using this procedure we then calculate the probability

(5.5.1) $P_b = \Pr\{\text{significance of both } y_1 \text{ and } y_2 \text{ in two steps}\}$.

As pointed out in McMillan and David (1971), the Murphy's procedure having a significance level $(\alpha+\alpha^2)/2$ has to be compared with the sequential procedure having a significance level α, so that under H_o the expected number of observations declared as outliers is equal for both the procedures. McMillan (1971) and Moran and McMillan (1973) have compared this measure P_b with the measure $P^{1,2}$ defined in last section. In addition to some other values, they have tabulated the values of P_b for $n = 21$, $\alpha = 0.05$ and $\theta_1 = \theta_2$ case. Methods for evaluation of $P^{1,2}$ by exact, approximate and Monte Carlo techniques are discussed in the last section. These are tabulated in Table 5.5.1 for $n = 21$, $\alpha = 0.02625$, where Monte Carlo values are based on 2000 iterations. It is worth mentioning that in Table 5.4.1 of last section, we have used exact critical value for Murphy's test.

But now we have only a nominal percentage point u_α. Consequently the values of $P^{1,2}$ as tabulated provide only a lower bound for exact values. Of course, since n and α are small, hence the difference is only marginal.

Since for $P^{1,2}$, we have used nominal percentage points, hence for a proper comparison, we also use nominal percentage points for the evaluation of P_b instead of exact percentage points. Nominal percentage points can be obtained from tables prepared by Srikantan (1961), Joshi (1975), Lund (1975), Doornbos (1980), and others. They can also be obtained directly by the methods described in these papers. We here point out that for n = 20, 21 and α = 0.05, the nominal percentage points obtained from Doornbos agree perfectly with the exact values tabulated by Grubbs and Beck (1972). The method given in Moran and McMillan (1973) for the evaluation of P_b by numerical integration is cumbersome. For our purposes we evaluate P_b for α = 0.05 and n = 21 by using Monte Carlo method, basing our results on 1000 iterations. These values are also tabulated in Table 5.5.1.

The simulated values of P_b using nominal percentage points agree reasonably well with the values of P_b tabulated by Moran and McMillan (1973) using exact percentage points. For example, for $\theta_1 = \theta_2 = 4$, the exact value of P_b is 0.421, while that obtained by simulation is 0.419. Similarly for $\theta_1 = \theta_2 = 5$, P_b (exact) = 0.759 and P_b (simulated) = 0.746.

From Table 5.5.1 we see that the Murphy's test perform much better than the sequential procedure, even if θ_1 and θ_2 are not equal. McMillan (1971) has come to the same conclusion for $\theta_1 = \theta_2$ case.

Corresponding results for the V statistic, which in this case is internally Studentized range statistic, for n = 20 and significance level 0.05 are tabled in Table 5.5.2. We now tabulate $P^{*1,2}$ for $\alpha = 0.02625$ by approximate method only, while P_b values for $\alpha = 0.05$ are obtained by simulation method with 1000 iterations.

For this case also, the V statistic performs much better than the sequential procedures for all values of θ_1 and θ_2 considered. In fact, the difference between $P^{*1,2}$ and P_b is substantial even if $\theta_1 \neq \theta_2$.

We next consider a two-way layout with r rows and c columns. For simplicity we shall consider the case $r < c$ in detail. Now from Theorem 5.4.1, we see that $P_a = P^{11,21}$, that is, the minimum value of $P^{i_1 j_1, i_2 j_2}$ occurs when the two outliers are in the same column. The values of $P^{11,21}$ for $\alpha = 0.02625$ for 4x5 and 6x8 tables are tabulated in Table 5.5.3 for $\theta_1 = \theta_2 = \theta$ case. The values for $\theta \leq 5$ are obtained by exact method, while for larger values of θ, these are obtained by approximate method.

For comparison, we also tabulate P_b values obtained by Monte Carlo techniques. Now θ is added to observed values of cells (1,1) and (2,1) and the data is tested for an outlier.

After one outlier is detected, we remove that observation and reanalyse the data for a second outlier. Again the nominal percentage points obtained from Doornbos (1980) can be used. In this manner we evaluate P_b for $\alpha = 0.05$ for 4x5 and 6x8 tables by simulation using 2000 iterations.

From Table 5.5.3 we see that the performance of the U statistic is remarkably higher than that of the sequential procedure, especially for 4x5 table, for which the sequential procedure performs very poorly.

For the statistic V, we see that the minimum values of $P^{*i_1 j_1, i_2 j_2}$ occurs when the outlying observations are neither in the same row nor in the same column, that is $P_a^* = P^{*11,22}$. These values of $P^{*11,22}$ are tabulated in Table 5.5.4 for $\alpha = 0.02625$ for 4x5 and 6x8 tables and for $\theta_1 = \theta_2 = \theta$. These are obtained by approximate method as described in the calculations for Table 5.4.7 of last section.

Again for comparison, we tabulate P_b values obtained by Monte Carlo method. In this case θ is subtracted from the observed value of cell (1,1) and θ is added to the observed value of cell (2,2). Computations are similar to that of U statistic.

Here again, we see from Table 5.5.4 that the statistic V performs considerably better than the sequential procedure, especially for 4x5 table.

Thus in all cases considered our proposed statistics U and V for two outliers perform better than the sequential test statistic.

TABLE 5.3.1. Exact and approximate values of $p^{*1,2}$ for a random sample of size n = 10 and $\alpha = 0.05$.

θ_1	θ_2	$p^{*1,2}$ exact	$p^{*1,2}$ approximate	Relative error
0	0	0.00111	0.00111	0.000
0	1	0.00269	0.00273	-0.015
0	2	0.00734	0.00737	-0.004
0	3	0.01341	0.01284	0.043
0	4	0.01789	0.01640	0.083
0	5	0.01915	0.01702	0.111
1	1	0.01258	0.01274	-0.013
1	2	0.03571	0.03545	0.007
1	3	0.06683	0.06548	0.020
1	4	0.09230	0.08957	0.030
1	5	0.10394	0.10022	0.036
2	2	0.10266	0.10089	0.017
2	3	0.19425	0.19137	0.015
2	4	0.27274	0.26973	0.011
2	5	0.31629	0.31387	0.008
3	3	0.36765	0.36443	0.009
3	4	0.51476	0.51285	0.004
3	5	0.60011	0.59979	0.001
4	4	0.70998	0.71031	0.000

TABLE 5.3.2. Exact and approximate values of $P^{1,2}$ for a random sample of size n = 10 and α = 0.05.

θ_1	θ_2	$P^{1,2}$ exact	$P^{1,2}$ approximate	Relative error
0	0	0.00111	0.00222	-1.000
0	1	0.00416	0.00435	-0.046
0	2	0.00925	0.00918	0.008
0	3	0.01360	0.01294	0.049
0	4	0.01478	0.01350	0.087
0	5	0.01297	0.01146	0.116
1	1	0.01744	0.01747	-0.002
1	2	0.04300	0.04236	0.015
1	3	0.06994	0.06814	0.026
1	4	0.08437	0.08135	0.036
1	5	0.08319	0.07949	0.044
2	2	0.11575	0.11312	0.023
2	3	0.20313	0.19929	0.019
2	4	0.26402	0.26015	0.015
2	5	0.28339	0.28002	0.012
3	3	0.37602	0.37182	0.011
3	4	0.50979	0.50723	0.005
3	5	0.57368	0.57278	0.002
4	4	0.70405	0.70437	0.000

TABLE 5.4.1. $P^{1,2}, P^{1,3}, P^{2,3}, P^{3,4}$ and min $(\bar{Q}_{12}, 1)$ values for $n = 10$ and $\alpha = 0.05$ obtained by exact method.

θ_1	θ_2	$P^{1,2}$	$P^{1,3}$	$P^{2,3}$	$P^{3,4}$	Min $(\bar{Q}_{12}, 1)$
0	0	.0011	.0011	.0011	.0011	.0500
0	1	.0042	.0012	.0042	.0012	.0807
0	2	.0093	.0007	.0093	.0007	.1082
0	3	.0136	.0002	.0136	.0002	.1304
0	4	.0148	.0000	.0148	.0000	.1344
0	5	.0130	.0000	.0130	.0000	.1169
1	1	.0174	.0022	.0022	.0014	.0914
1	2	.0430	.0006	.0051	.0009	.1136
1	3	.0699	.0001	.0080	.0003	.1430
1	4	.0844	.0000	.0090	.0001	.1581
1	5	.0832	.0000	.0081	.0000	.1482
2	2	.1158	.0015	.0015	.0006	.1579
2	3	.2031	.0003	.0027	.0003	.2332
2	4	.2640	.0000	.0033	.0001	.2919
2	5	.2834	.0000	.0031	.0000	.3087
3	3	.3760	.0005	.0005	.0001	.3868
3	4	.5098	.0001	.0007	.0000	.5163
3	5	.5737	.0000	.0007	.0000	.5796
4	4	.7041	.0001	.0001	.0000	.7055
4	5	.8031	.0000	.0001	.0000	.8039
5	5	.9106	.0000	.0000	.0000	.9107

TABLE 5.4.2. Approximate upper limit (top row) and lower limit (bottom row) for the power of Murphy's test for $n = 10$, $\alpha = 0.05$.

θ_2 \ θ_1	0	1	2	3	4	5	6	7
0	0.100							
	0.002							
1	0.101	0.100						
	0.004	0.017						
2	0.109	0.111	0.150					
	0.009	0.042	0.113					
3	0.122	0.135	0.224	0.380				
	0.013	0.068	0.199	0.372				
4	0.122	0.147	0.283	0.512	0.705			
	0.014	0.081	0.260	0.507	0.704			
5	0.103	0.136	0.301	0.577	0.805	0.916		
	0.011	0.079	0.280	0.573	0.805	0.916		
6	0.076	0.109	0.283	0.587	0.839	0.953	0.986	
	0.008	0.068	0.267	0.583	0.838	0.953	0.986	
7	0.049	0.079	0.245	0.560	0.837	0.962	0.993	0.999
	0.005	0.052	0.234	0.558	0.837	0.962	0.993	0.999

TABLE 5.4.3. $P^{1,2}, P^{1,3}, P^{2,3}, P^{3,4}$ and Q_{12} values for n = 10 and α = 0.05 obtained by Monte Carlo method

θ_1	θ_2	$P^{1,2}$	$P^{1,3}$	$P^{2,3}$	$P^{3,4}$	Q_{12}
0	0	.0011	.0013	.0012	.0013	.0519
0	1	.0038	.0005	.0041	.0005	.0599
0	2	.0102	.0001	.0100	.0001	.0946
0	3	.0145	.0000	.0140	.0000	.1275
0	4	.0161	.0000	.0149	.0000	.1351
0	5	.0145	.0000	.0128	.0000	.1170
1	1	.0180	.0023	.0024	.0003	.0640
1	2	.0449	.0007	.0056	.0001	.0972
1	3	.0746	.0001	.0083	.0000	.1426
1	4	.0882	.0000	.0092	.0000	.1619
1	5	.0852	.0000	.0083	.0000	.1513
2	2	.1156	.0016	.0018	.0000	.1433
2	3	.2082	.0004	.0028	.0000	.2337
2	4	.2653	.0000	.0033	.0000	.2922
2	5	.2846	.0000	.0032	.0000	.3100
3	3	.3761	.0006	.0005	.0000	.3847
3	4	.5103	.0001	.0007	.0000	.5171
3	5	.5753	.0000	.0007	.0000	.5808
4	4	.7048	.0002	.0000	.0000	.7066
4	5	.8029	.0000	.0001	.0000	.8033
5	5	.9124	.0000	.0000	.0000	.9126

TABLE 5.4.4. $p^{*1,2}, p^{*1,3}, p^{*2,3}, p^{*3,4}$ and min $(\bar{Q}^{*}_{12}, 1)$ values for n = 10 and α = 0.05 obtained by exact method.

θ_1	θ_2	$p^{*1,2}$	$p^{*1,3}$	$p^{*2,3}$	$p^{*3,4}$	Min $(\bar{Q}^{*}_{12}, 1)$
0	0	.0011	.0011	.0011	.0011	.0499
0	1	.0027	.0008	.0027	.0008	.0524
0	2	.0073	.0003	.0073	.0003	.0556
0	3	.0134	.0001	.0134	.0001	.1224
0	4	.0179	.0000	.0179	.0000	.1612
0	5	.0192	.0000	.0192	.0000	.1724
1	1	.0126	.0018	.0018	.0005	.0556
1	2	.0357	.0006	.0047	.0002	.0831
1	3	.0668	.0001	.0084	.0000	.1360
1	4	.0923	.0000	.0111	.0000	.1814
1	5	.1039	.0000	.0118	.0000	.1984
2	2	.1027	.0017	.0017	.0001	.1305
2	3	.1943	.0003	.0030	.0000	.2212
2	4	.2727	.0000	.0041	.0000	.3059
2	5	.3163	.0000	.0045	.0000	.3524
3	3	.3677	.0006	.0006	.0000	.3775
3	4	.5148	.0001	.0009	.0000	.5224
3	5	.6001	.0000	.0010	.0000	.6085
4	4	.7100	.0001	.0001	.0000	.7117
4	5	.8180	.0000	.0001	.0000	.8187
5	5	.9232	.0000	.0000	.0000	.9233

TABLE 5.4.5. Approximate upper limit (top row) and lower limit (bottom row) for the power of Studentized range test for n = 10, $\alpha = 0.05$.

θ_2 \ θ_1	0	1	2	3	4	5	6	7
0	0.050							
	0.001							
1	0.052	0.054						
	0.003	0.013						
2	0.074	0.080	0.125					
	0.007	0.035	0.101					
3	0.116	0.129	0.214	0.371				
	0.013	0.065	0.191	0.364				
4	0.148	0.168	0.297	0.518	0.711			
	0.016	0.090	0.270	0.513	0.710			
5	0.153	0.181	0.343	0.606	0.819	0.923		
	0.017	0.100	0.314	0.600	0.818	0.923		
6	0.138	0.171	0.351	0.640	0.863	0.961	0.989	
	0.015	0.098	0.325	0.635	0.862	0.961	0.989	
7	0.112	0.146	0.334	0.640	0.875	0.972	0.995	0.999
	0.012	0.088	0.313	0.635	0.874	0.971	0.995	0.999

TABLE 5.4.6. $p^{*1,2}, p^{*1,3}, p^{*2,3}, p^{*3,4}$ and Q_{12}^{*} values for n = 10 and α = 0.05 obtained by Monte Carlo method

θ_1	θ_2	$p^{*1,2}$	$p^{*1,3}$	$p^{*2,3}$	$p^{*3,4}$	Q_{12}^{*}
0	0	.0009	.0011	.0011	.0011	.0494
0	1	.0035	.0009	.0029	.0008	.0553
0	2	.0089	.0002	.0078	.0003	.0804
0	3	.0138	.0001	.0140	.0000	.1275
0	4	.0184	.0000	.0183	.0000	.1646
0	5	.0208	.0000	.0193	.0000	.1751
1	1	.0130	.0016	.0017	.0005	.0526
1	2	.0359	.0006	.0050	.0002	.0850
1	3	.0670	.0001	.0089	.0000	.1396
1	4	.0970	.0000	.0117	.0000	.1903
1	5	.1100	.0000	.0123	.0000	.2081
2	2	.0998	.0016	.0017	.0000	.1268
2	3	.1989	.0003	.0032	.0000	.2265
2	4	.2757	.0000	.0042	.0000	.3091
2	5	.3168	.0000	.0047	.0000	.3542
3	3	.3770	.0006	.0006	.0000	.3865
3	4	.5179	.0000	.0009	.0000	.5255
3	5	.5994	.0000	.0011	.0000	.6079
4	4	.7121	.0001	.0001	.0000	.7134
4	5	.8213	.0000	.0001	.0000	.8224
5	5	.9237	.0000	.0000	.0000	.9239

TABLE 5.4.7. Approximate values of $P_a^* = P^{*11,22}$ for two-way 4x5 (top row) and 6x8 tables (bottom row), for $\alpha = 0.02625$.

θ_2 \ θ_1	0	1	2	3	4	5	6	7
0	0.000							
	0.000							
1	0.000	0.001						
	0.000	0.001						
2	0.001	0.004	0.011					
	0.001	0.005	0.018					
3	0.002	0.008	0.025	0.057				
	0.003	0.015	0.051	0.130				
4	0.002	0.012	0.041	0.101	0.188			
	0.008	0.037	0.114	0.261	0.459			
5	0.002	0.014	0.055	0.144	0.278	0.421		
	0.019	0.075	0.209	0.423	0.653	0.828		
6	0.002	0.015	0.062	0.173	0.347	0.534	0.682	
	0.036	0.131	0.326	0.583	0.802	0.927	0.978	
7	0.002	0.013	0.061	0.184	0.384	0.603	0.773	0.872
	0.062	0.202	0.449	0.715	0.894	0.972	0.994	0.999

TABLE 5.5.1. Exact, approximate and simulated values of $P^{1,2}$ for the statistic U with $\alpha = 0.02625$ and simulated values of P_b for the sequential test with $\alpha = 0.05$ and $n = 21$.

θ_1	θ_2	$P^{1,2}$ exact	$P^{1,2}$ approximate	$P^{1,2}$ simulated	P_b simulated
0	0	0.000	0.000	0.000	0.000
0	1	0.001	0.001	0.001	0.000
0	2	0.004	0.004	0.002	0.001
0	3	0.010	0.011	0.012	0.003
0	4	0.022	0.023	0.019	0.001
0	5	0.036	0.036	0.035	0.002
1	1	0.006	0.006	0.005	0.002
1	2	0.022	0.023	0.023	0.000
1	3	0.057	0.057	0.056	0.012
1	4	0.109	0.107	0.105	0.020
1	5	0.167	0.164	0.152	0.020
2	2	0.078	0.077	0.073	0.017
2	3	0.182	0.177	0.180	0.046
2	4	0.311	0.304	0.317	0.094
2	5	0.433	0.427	0.430	0.119
3	3	0.376	0.368	0.377	0.132
3	4	0.572	0.567	0.578	0.244
3	5	0.718	0.718	0.721	0.342
4	4	0.779	0.781	0.771	0.419
4	5	0.895	0.900	0.900	0.581
5	5	0.970	0.971	0.973	0.746

TABLE 5.5.2. Approximate value of $P^{*1,2}$ for the statistic V with $\alpha = 0.02625$ and simulated value of P_b for the sequential test with $\alpha = 0.05$ and $n = 20$.

θ_1	θ_2	$P^{*1,2}$ approximate	P_b simulated
0	0	0.000	0.000
0	1	0.001	0.000
0	2	0.003	0.000
0	3	0.009	0.000
0	4	0.019	0.002
0	5	0.032	0.002
1	1	0.005	0.000
1	2	0.018	0.002
1	3	0.048	0.002
1	4	0.094	0.011
1	5	0.150	0.017
2	2	0.065	0.006
2	3	0.157	0.015
2	4	0.281	0.049
2	5	0.405	0.088
3	3	0.342	0.074
3	4	0.542	0.159
3	5	0.699	0.251
4	4	0.763	0.317
4	5	0.889	0.492
5	5	0.967	0.680

TABLE 5.5.3. The $P^{11,21}$ values for the statistic U with $\alpha = 0.02625$ and simulated values of P_b for the sequential test with $\alpha = 0.05$ for two-way tables. The values shown with asterisks denote exact values.

θ	4x5 table		6x8 table	
	$P^{11,21}$	P_b	$P^{11,21}$	P_b
0	0.000*	0.000	0.000*	0.000
1	0.002*	0.001	0.001*	0.001
2	0.011*	0.003	0.015*	0.005
3	0.048*	0.004	0.113*	0.017
4	0.150*	0.005	0.395*	0.078
5	0.336*	0.005	0.747*	0.203
6	0.569	0.007	0.951	0.394
7	0.780	0.008	0.996	0.533
8	0.914	0.008	1.000	0.651
9	0.974	0.010	1.000	0.763

TABLE 5.5.4. The $p^{*11,22}$ values for the statistic V with $\alpha = 0.02625$ and simulated values of P_b for the sequential test with $\alpha = 0.05$ for two-way tables.

θ	4x5 table		6x8 table	
	$p^{*11,22}$	P_b	$p^{*11,22}$	P_b
0	0.000	0.000	0.000	0.000
1	0.001	0.000	0.001	0.000
2	0.011	0.000	0.018	0.001
3	0.057	0.003	0.130	0.012
4	0.188	0.005	0.459	0.099
5	0.421	0.014	0.828	0.288
6	0.682	0.026	0.978	0.533
7	0.872	0.035	0.999	0.720
8	0.963	0.040	1.000	0.836
9	0.992	0.039	1.000	0.910

CHAPTER VI

EXTENSION FOR MORE THAN TWO OUTLIERS

6.1. Introduction

In previous chapters we were dealing with the case when only two outliers were present. For three or more outliers, earlier work has been done by Rosner (1975), Draper and John (1980), Gentleman (1980) etc. Bradu and Hawkins (1982) has also discussed this multiple outlier case using tetrads.

We now extend our results for three or more outliers for one sided case, that is, when all outliers are in the same direction. The statistic is introduced in Section 6.2 and distribution theory results are given in Section 6.3. Nominal upper percentage points have been tabulated for three-outlier case in Section 6.4. The performance of test statistic is studied in Section 6.5.

6.2. Motivation of the statistic

The u_{ij}'s for the one-sided statistic U in two-outlier case are defined by

$$u_{ij} = t_{ij}/S_p,$$

where $$t_{ij} = S_p\,(w_i+w_j)/[2(1+\rho_{ij})]^{1/2}$$

$$= (e_i/\lambda_{ii}^{1/2} + e_j/\lambda_{jj}^{1/2})/[2(1+\rho_{ij})]^{1/2}$$

has the variance equal to σ^2.

When three outliers are present on the right, we define the test statistic for detecting them as follows :

Let

(6.2.1) $t_{hij} = C_{hij}\, S_p(w_h + w_i + w_j)$, and

(6.2.2) $u_{hij} = t_{hij}/S_p = C_{hij}(w_h + w_i + w_j)$,

where w_i and S_p are defined in Section 2.1. We now choose C_{hij} so that the variance of t_{hij} is σ^2, and define the statistic $U^{(3)}$ for three outliers by

$$U^{(3)} = \underset{1\le h<i<j\le n}{\text{Max}}\ u_{hij}.$$

Now,

$$\text{Var}\,(t_{hij}) = \text{Var}\,[C_{hij}(e_h/\lambda_{hh}^{1/2}+e_i/\lambda_{ii}^{1/2}+e_j/\lambda_{jj}^{1/2})]$$

$$= C_{hij}^2\sigma^2(1+1+1+2\rho_{hi} + 2\rho_{hj} + 2\rho_{ij}).$$

Equating it to σ^2 we immediately get for finite C_{hij},

$$C_{hij} = 1/[3+2(\rho_{hi} + \rho_{hj} + \rho_{ij})]^{1/2}$$

and

$$u_{hij} = (w_h+w_i+w_j)/[3+2(\rho_{hi} + \rho_{hj} + \rho_{ij})]^{1/2}.$$

For a random sample of size n from $N(\mu,\sigma^2)$, we have $\rho_{ij} = \rho = -1/(n-1)$ for all i and j and $\lambda_{ii} = (n-1)/n$, for all i. Hence for $\nu = 0$, we have

$$C_{hij} = [(n-1)/\{3(n-3)\}]^{1/2},$$

and $$u_{hij} = K(y_h + y_i + y_j - 3\bar{y})/s,$$

where $K = [n/\{3(n-1)(n-3)\}]^{1/2}$.

This gives

$$(6.2.3) \quad U^{(3)} = K\,[y_{(n)} + y_{(n-1)} + y_{(n-2)} - 3\bar{y}]/s = K\,T_{N3},$$

where T_{N3} is the Murphy's statistic for three outliers.

Generalization to m_1 outliers on the right are now immediate. The statistic $U^{(m_1)}$ is now defined by

$$U^{(m_1)} = \max_{1 \le i_1 < i_2 < \ldots < i_{m_1} \le n} u_{i_1, i_2, \ldots, i_{m_1}},$$

where

$$u_{i_1, i_2, \ldots, i_{m_1}} = \frac{w_{i_1} + w_{i_2} + \ldots + w_{i_{m_1}}}{[m_1 + \sum\sum_{g \neq h = 1} \rho_{i_g i_h}]^{1/2}}.$$

Similar to the case of three outliers, it is equivalent to Murphy's test for m_1 outliers for a random sample from $N(\mu, \sigma^2)$ distribution. The statistic U for two outliers can thus be extended for m_1 outliers, but statistic V cannot be extended in this manner.

6.3. Distribution theory

We shall discuss the distribution theory of these u's for the case of three outliers in detail. Results for the general case are analogous.

From equation (6.2.2) we have

$$u_{hij} = c_{hij}\,(w_h + w_i + w_j), \quad 1 \le h < i < j \le n,$$

where

$$(6.3.1) \qquad c_{hij} = 1/[3+2(\rho_{hi} + \rho_{hj} + \rho_{ij})]^{1/2}.$$

Now applying Corollary 2.2.1 with

$$M = c_{hij}\,(0,0,\ldots,0,1,0,\ldots,0,1,\ldots,0,\ldots,1,0,\ldots,0),$$

where 1 occurs in h,i and jth places, and noting that $C = \underset{\sim}{M}\,\underset{\sim}{R}\,\underset{\sim}{M}' = 1$, we get the marginal pdf of u_{hij} as

$$(6.3.2) \quad f(u_{hij}) = (1-u_{hij}^2)^{(p-3)/2}/B[1/2,(p-1)/2], -1 \le u_{hij} \le 1,$$

where $p = n-k+\nu$.

It is clear that for general m_1 the marginal distribution of $u_{i_1,i_2,\ldots,i_{m_1}}$ will be identical as that of u_{hij} given at equation (6.3.2). Other joint distributions can be obtained from Theorem 2.2.1 in a like manner.

6.4. Percentile points

Nominal upper percentile point $u_\alpha^{(3)}$ for $U^{(3)}$ can be obtained by using first Bonferroni inequality. We have

$$U^{(3)} = \max_{1\le h<i<j\le n} u_{hij}.$$

Hence,

$$\Pr(U^{(3)} > u_\alpha^{(3)}) \le \sum_{1\le h<i<j\le n} \Pr(u_{hij} \ge u_\alpha^{(3)})$$

$$= \binom{n}{3} \Pr(u_{hij} \ge u_\alpha^{(3)}),$$

since marginal distribution of each u_{hij} is the same. Consequently

a nominal upper 100α percentage point $u_\alpha^{(3)}$ is given by

$$(6.4.1) \quad \Pr(u_{hij} \geq u_\alpha^{(3)}) = \alpha/\binom{n}{3} = 6\alpha/[n(n-1)(n-2)] \ .$$

Or, equivalently from Lemma 2.3.1 for $u_\alpha^{(3)} \geq 0$, it is given by

$$(6.4.2) \quad I_{1-u_\alpha^{(3)}}{}^2 [(p-1)/2, 1/2] = 12\alpha/[n(n-1)(n-2)] \ .$$

Solving equation (6.4.2) for $u_\alpha^{(3)}$, a nominal upper percentile point for $U^{(3)}$ is obtained.

Similarly for m_1 outliers a nominal upper 100α percentage point $u_\alpha^{(m_1)}$ of $U^{(m_1)}$ is given by

$$(6.4.3) \quad I_{1-u_\alpha^{(m_1)}}{}^2 [(p-1)/2, 1/2] = 2\alpha/\binom{n}{m_1} .$$

These nominal upper percentile points $u_\alpha^{(3)}$ for $\alpha = 0.01, 0.05$ are given in Table 6.4.1. Similar to Table 2.3.1 these are tabulated for $\nu = 0$, $n = 5(1)12, 14(1)16(2)20, 21, 24, 25, 27, 28(2)32, 33, 35, 36, 40, 42, 45, 48(1)50(5)60(10)100$ and $k = 1(1) \min(15, n-2)$.

Using the relation given at equation (6.2.3), nominal percentiles for T_{N3} statistic are obtained for $n = 10, 20, 50$ and 100. These are given in Table 6.4.2. These values can be compared with the simulated critical points of T_{N3} obtained by Barnett and Lewis (1978). From this we find that our values agree with their values considerably for $n \leq 50$.

6.5. Performance of the statistics

We now study the performance of test statistic proposed in Section 6.2 in the non-null situation when three outliers are present on the right. The distribution theory results for this case are direct generalizations of results for two-outlier case. Further, for simplicity we shall consider the case $\nu = 0$ in detail. The null hypothesis specifies that there are no outliers present, while the alternative hypothesis is the union of $\binom{n}{3}$ hypotheses H_{hij} $(1 \leq h < i < j \leq n)$.

6.5.1. Distribution Theory

Without loss of generality, we consider $(h,i,j) = (1,2,3)$. Then under H_{123},

$$\underset{\sim}{y} \overset{d}{=} N(\cdot, \sigma^2 \underset{\sim}{I}),$$

$$E(\underset{\sim}{y}) = \underset{\sim}{X}\,\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}_1\theta_1 + \underset{\sim}{\varepsilon}_2\theta_2 + \underset{\sim}{\varepsilon}_3\theta_3 ,$$

where $\underset{\sim}{\varepsilon}_s (s = 1,2,\ldots,n)$, is the sth column of $\underset{\sim}{I}_n$. Further $\theta_i > 0$ for $i = 1,2,3$.

Under H_{123}, the residual vector, $\underset{\sim}{e} = \underset{\sim}{\Lambda}\,\underset{\sim}{y}$ is normally distributed with variance covariance matrix $\underset{\sim}{\Lambda}\sigma^2$ and mean

$$\begin{aligned} E(\underset{\sim}{e}|H_{123}) &= \underset{\sim}{\Lambda}\, E(\underset{\sim}{y}|H_{123}) \\ &= \underset{\sim}{\Lambda}\,[E(\underset{\sim}{y}|H_o) + \underset{\sim}{\varepsilon}_1\theta_1 + \underset{\sim}{\varepsilon}_2\theta_2 + \underset{\sim}{\varepsilon}_3\theta_3] \\ &= \underset{\sim}{\Lambda}\,\underset{\sim}{X}\,\underset{\sim}{\beta} + \underset{\sim}{\Lambda}\,(\underset{\sim}{\varepsilon}_1\theta_1 + \underset{\sim}{\varepsilon}_2\theta_2 + \underset{\sim}{\varepsilon}_3\theta_3) \\ &= \underset{\sim}{\Lambda}\,(\underset{\sim}{\varepsilon}_1\theta_1 + \underset{\sim}{\varepsilon}_2\theta_2 + \underset{\sim}{\varepsilon}_3\theta_3), \end{aligned}$$

since $\underset{\sim}{\Lambda}_2 \underset{\sim}{X}_2 = \underset{\sim}{0}_2$. The error sum of squares S^2 has a non-central $\sigma^2\chi^2$ distribution with n-k degrees of freedom and non-centrality parameter λ^*, where

$$\sigma^2\lambda^* = E(\underset{\sim}{y}'_2 | H_{123}) \underset{\sim}{\Lambda}_2 E(\underset{\sim}{y}_2 | H_{123})$$

$$= (\underset{\sim}{\beta}'_2 \underset{\sim}{X}'_2 + \underset{\sim}{\varepsilon}'_1\theta_1 + \underset{\sim}{\varepsilon}'_2\theta_2 + \underset{\sim}{\varepsilon}'_3\theta_3) \underset{\sim}{\Lambda}_2 (\underset{\sim}{X}_2 \underset{\sim}{\beta}_2 + \underset{\sim}{\varepsilon}_1\theta_1 + \underset{\sim}{\varepsilon}_2\theta_2 + \underset{\sim}{\varepsilon}_3\theta_3)$$

$$= \sum_{i=1}^{3} \sum_{j=1}^{3} \underset{\sim}{\varepsilon}'_i \underset{\sim}{\Lambda}_2 \underset{\sim}{\varepsilon}_j \theta_i\theta_j$$

$$(6.5.1) = \lambda_{11}\theta_1^2 + \lambda_{22}\theta_2^2 + \lambda_{33}\theta_3^2 + 2\lambda_{12}\theta_1\theta_2 + 2\lambda_{13}\theta_1\theta_3 + 2\lambda_{23}\theta_2\theta_3.$$

From equation (6.2.1), we have

$$t_{hij} = c_{hij}(e_h/\lambda_{hh}^{1/2} + e_i/\lambda_{ii}^{1/2} + \varepsilon_j/\lambda_{jj}^{1/2})$$

$$= c_{hij}(\underset{\sim}{\lambda}'_{(h)}/\lambda_{hh}^{1/2} + \underset{\sim}{\lambda}'_{(i)}/\lambda_{ii}^{1/2} + \underset{\sim}{\lambda}'_{(j)}/\lambda_{jj}^{1/2}) \underset{\sim}{y}_2$$

$$= \underset{\sim}{c}'_2 \underset{\sim}{y}_2 \text{ (say)},$$

where $\underset{\sim}{\lambda}_{(i)}$ is the ith column of $\underset{\sim}{\Lambda}_2$, c_{hij} is given at equation (6.3.1) and

$$\underset{\sim}{c}'_2 = c_{hij}(\underset{\sim}{\lambda}'_{(h)}/\lambda_{hh}^{1/2} + \underset{\sim}{\lambda}'_{(i)}/\lambda_{ii}^{1/2} + \underset{\sim}{\lambda}'_{(j)}/\lambda_{jj}^{1/2}).$$

Note that $\underset{\sim}{c}'_2 \underset{\sim}{c}_2 = 1$, and variance of t_{hij} is σ^2. Further

$$\underset{\sim}{\Lambda}_2 \underset{\sim}{c}_2 = \begin{bmatrix} \underset{\sim}{\lambda}'_{(1)} \\ \underset{\sim}{\lambda}'_{(2)} \\ \vdots \\ \underset{\sim}{\lambda}'_{(n)} \end{bmatrix} c_{hij}(\underset{\sim}{\lambda}_{(h)}/\lambda_{hh}^{1/2} + \underset{\sim}{\lambda}_{(i)}/\lambda_{ii}^{1/2} + \underset{\sim}{\lambda}_{(j)}/\lambda_{jj}^{1/2})$$

$$= c_{hij}\begin{bmatrix} \lambda_{1h}/\lambda_{hh}^{1/2} + \lambda_{1i}/\lambda_{ii}^{1/2} + \lambda_{1j}/\lambda_{jj}^{1/2} \\ \lambda_{2h}/\lambda_{hh}^{1/2} + \lambda_{2i}/\lambda_{ii}^{1/2} + \lambda_{2j}/\lambda_{jj}^{1/2} \\ \vdots \\ \lambda_{nh}/\lambda_{hh}^{1/2} + \lambda_{ni}/\lambda_{ii}^{1/2} + \lambda_{nj}/\lambda_{jj}^{1/2} \end{bmatrix}$$

$$= c_{hij}(\underset{\sim}{\lambda}_{(h)}/\lambda_{hh}^{1/2} + \underset{\sim}{\lambda}_{(i)}/\lambda_{ii}^{1/2} + \underset{\sim}{\lambda}_{(j)}/\lambda_{jj}^{1/2})$$

$$= \underset{\sim}{c}.$$

Since t_{hij} is a linear combination of $\underset{\sim}{y}$, hence under H_{123},

$$t_{hij} \overset{d}{=} N(\mu,\sigma^2),$$

where

$$\mu = E(t_{hij}|H_{123}) = \underset{\sim}{c}' E(\underset{\sim}{y}|H_{123})$$

$$= c_{hij}\,[(\lambda_{1h}/\lambda_{hh}^{1/2} + \lambda_{1i}/\lambda_{ii}^{1/2} + \lambda_{1j}/\lambda_{jj}^{1/2})\theta_1$$

$$+ (\lambda_{2h}/\lambda_{hh}^{1/2} + \lambda_{2i}/\lambda_{ii}^{1/2} + \lambda_{2j}/\lambda_{jj}^{1/2})\,\theta_2$$

$$+ (\lambda_{3h}/\lambda_{hh}^{1/2} + \lambda_{3i}/\lambda_{ii}^{1/2} + \lambda_{3j}/\lambda_{jj}^{1/2})\,\theta_3]$$

(6.5.2) $$= c_{hij}\,[(\rho_{1h} + \rho_{1i} + \rho_{1j})\,\lambda_{11}^{1/2}\theta_1 + (\rho_{2h}+\rho_{2i}+\rho_{2j})\lambda_{22}^{1/2}\theta_2$$

$$+ (\rho_{3h}+\rho_{3i}+\rho_{3j})\,\lambda_{33}^{1/2}\theta_3]\,.$$

Now, if we let $Q_1 = t_{hij}^2$, and $Q_2 = S^2 - t_{hij}^2$, then

$$Q_1 \overset{d}{=} \sigma^2\,\chi^2(1,\delta),$$

where the non-centrality parameter δ is given by

(6.5.3) $$\sigma^2\delta = E(\underset{\sim}{y}'|H_{123})\;\underset{\sim}{c}\;\underset{\sim}{c}'\;E(\underset{\sim}{y}|H_{123}) = \mu^2,$$

and μ is given at equation (6.5.2).

Next, $$Q_2 = S^2 - t_{hij}^2$$

$$= \underset{2}{y}' \underset{2}{\Lambda} \underset{2}{y} - \underset{2}{y}' \underset{2}{c} \underset{2}{c}' \underset{2}{y}$$

$$= \underset{2}{y}' [\underset{2}{\Lambda} - \underset{2}{c} \underset{2}{c}'] \underset{2}{y}.$$

Now $$(\underset{2}{\Lambda} - \underset{2}{c} \underset{2}{c}')(\underset{2}{\Lambda} - \underset{2}{c} \underset{2}{c}') = \underset{2}{\Lambda} - \underset{2}{\Lambda} \underset{2}{c} \underset{2}{c}' - \underset{2}{c} \underset{2}{c}' \underset{2}{\Lambda} + \underset{2}{c} \underset{2}{c}' \underset{2}{c} \underset{2}{c}'$$

$$= \underset{2}{\Lambda} - \underset{2}{c} \underset{2}{c}' - \underset{2}{c} \underset{2}{c}' + \underset{2}{c} \underset{2}{c}' = \underset{2}{\Lambda} - \underset{2}{c} \underset{2}{c}'.$$

Thus $\underset{2}{\Lambda} - \underset{2}{c} \underset{2}{c}'$ is an idempotent matrix. Further

$$\text{rank}\,(\underset{2}{\Lambda} - \underset{2}{c} \underset{2}{c}') = \text{tr}\,(\underset{2}{\Lambda} - \underset{2}{c} \underset{2}{c}')$$

$$= \text{tr}\,(\underset{2}{\Lambda}) - \text{tr}\,(\underset{2}{c} \underset{2}{c}') = n-k-1.$$

Consequently, $Q_2 \overset{d}{=} \sigma^2 \chi^2(n-k-1, \eta)$, where the non-centrality parameter η is given by

$$\sigma^2 \eta = E(\underset{2}{y}' | H_{123})(\underset{2}{\Lambda} - \underset{2}{c} \underset{2}{c}') E(\underset{2}{y} | H_{123}) \qquad (6.5.4)$$

$$= \sigma^2 \lambda^* - \sigma^2 \delta,$$

where $\sigma^2 \lambda^*$ and $\sigma^2 \delta$ are given at equations (6.5.1) and (6.5.3) respectively.

Finally t_{hij} and Q_2 are independent since

$$(\underset{2}{\Lambda} - \underset{2}{c} \underset{2}{c}')\underset{2}{c} = \underset{2}{\Lambda} \underset{2}{c} - \underset{2}{c} \underset{2}{c}' \underset{2}{c} = \underset{2}{c} - \underset{2}{c} = \underset{2}{0}.$$

We thus have a result analogous to Lemma 5.2.1. The distribution of

$$u_{hij} = t_{hij} / (t_{hij}^2 + Q_2)^{1/2}$$

is now obtained from Theorem 5.2.1, with μ, δ and η given by

equations (6.5.2), (6.5.3) and (6.5.4) respectively. This is given (with $u = u_{hij}$, $\mu > 0$, etc.) by

$$f(u) = \sum_{j=0}^{\infty} K_j \sum_{i=0}^{\infty} K_i^* u^i (1-u^2)^{j+a/2-1} / B[(i+1)/2, j+a/2], \quad -1 \le u \le 1,$$

where

$$K_j = e^{-\eta/2} (\eta/2)^j / j! \quad , \; j = 0,1,2,\ldots,$$

$$K_i^* = e^{-\delta/2} (\delta/2)^{i/2} / G(i/2+1), \; i = 0,1,2,\ldots,$$

$$a = n-k-1 \text{ and } \delta = \mu^2/\sigma^2.$$

Using this, $\Pr(u_{hij} > u_\alpha^{(3)} | H_{123})$ etc. can now be evaluated exactly as well as approximately by applying the techniques discussed in Sections 5.2 and 5.3.

6.5.2. Measures of performance

Similar to the case of two outliers, we consider the following measures of performance.

$$P_{hij} = \Pr(u_{hij} > u_\alpha^{(3)} | H_{hij}),$$

$$Q_{hij} = \Pr(U^{(3)} > u_\alpha^{(3)} | H_{hij}),$$

$$P_a = \min_{1 \le h < i < j \le n} P_{hij},$$

and

$$Q_a = \min_{1 \le h < i < j \le n} Q_{hij}.$$

If P_{hij} does not depend on h,i and j, then $P_a = P_{123}$ is the probability that u_{123} is significantly large when H_{123} is true. Similarly if Q_{hij} does not depend on h,i and j, then $Q_a = Q_{123}$ is the power function.

For the remainder of this section, we shall confine our attention to the case of a random sample from a normal population. As shown in equation (6.2.3), now the test statistic $U^{(3)}$ is equivalent to Murphy's test statistic for 3 outliers. Further P_{hij} does not depend on h,i and j. Consequently, $P_a = P_{123}$. Further $\lambda_{ii} = (n-1)/n = \lambda$ and $\rho_{ij} = -1/(n-1) = \rho$ for $i \neq j$. The parameters involved in the evaluation of P_{123} are obtained from equations (6.5.3) and (6.5.4), with h = 1, i = 2 and j = 3 in these equations. These are given by

$$\sigma^2 \delta = \lambda[(1+2\rho)/3] (\theta_1+\theta_2+\theta_3)^2$$

$$= [(n-3)/3n] (\theta_1+\theta_2+\theta_3)^2$$

and

$$\sigma^2 \eta = \lambda[2(1-\rho)/3] (\theta_1^2+\theta_2^2+\theta_3^2 - \theta_1\theta_2 - \theta_1\theta_3 - \theta_2\theta_3)$$

$$= (2/3) (\theta_1^2+\theta_2^2+\theta_3^2 - \theta_1\theta_2 - \theta_1\theta_3 - \theta_2\theta_3).$$

The evaluation of P_{123} for Murphy's test, although straight forward, is rather time consuming for all combinations of θ_1, θ_2 and θ_3. We therefore consider the case $\theta_1 = \theta_2 = \theta_3 = \theta$, and tabulate P_{123} for $\alpha = 0.05$ and n = 20 and 50 in Tables 6.5.1 and 6.5.2 respectively. The values of P_{123} are calculated by exact formula for $\theta \leq 3$ and by approximate formula for $\theta \geq 4$. As can be seen from these tables, the performance is quite good if 3 outliers are really present. For the sake of comparison, we also tabulate P_{123}, when (i) $\theta_1 = \theta_2 = \theta$, $\theta_3 = 0$; and (ii) $\theta_1 = \theta$, $\theta_2 = \theta_3 = 0$ in last two columns. The first case corresponds to

the situation when only two outliers on the right are present and Murphy's test for 3 outliers is applied. In this case the performance is not very good especially for n = 20. However for n = 50, and $\theta \geq 6$, the test performs reasonably well.

The second case with $\theta_1 = \theta$ and $\theta_2 = \theta_3 = 0$, corresponds to the situation when there is only one outlier and a test for 3 outliers is applied. For this case the performance is extremely poor even for n = 50.

Next consider the measure $Q_a \equiv Q_{123}$. Clearly $P_{123} \leq Q_{123}$, that is $P_a \leq Q_a$. This gives a lower bound for Q_a. An upper bound for $Q_a = \Pr(U^{(3)} > u_\alpha^{(3)} | H_{123})$, can be obtained by applying the first Bonferroni inequality. Thus

$$Q_a \leq \text{Min } (1, \bar{Q}_{123}),$$

where

$$\bar{Q}_{123} = \sum_{1 \leq h < i < j \leq n} \Pr(u_{hij} > u_\alpha^{(3)} | H_{123}).$$

Since the number of terms to be added is now considerably larger than the number of terms for two-outlier case, hence this bound is not very useful, except for very small values of n. We therefore evaluate Q_a, by simulation using 1000 iterations for n = 20 and 50, $\alpha = 0.05$ and for the case $\theta_1 = \theta_2 = \theta_3 = \theta$. In Tables 6.5.1 and 6.5.2, we tabulate simulated values of P_a and Q_a in third and fourth columns respectively.

Note that the simulated values of P_a agree considerably with the theoretical values. The values of Q_a are close to P_a for larger values of θ. However, these are not so close for small values of θ, as P_a only provides a lower bound for Q_a.

TABLE 6.4.1. Nominal upper critical values $u_\alpha^{(3)}$ of one-sided test statistic $U^{(3)}$ for three outliers in linear regression

(α = 0.05)

n|k	1	2	3	4	5	6	7
5	.9587	.9900	.9999				
6	.9417	.9740	.9950	1.0000			
7	.9248	.9560	.9821	.9971	1.0000		
8	.9085	.9379	.9653	.9870	.9982	1.0000	
9	.8929	.9203	.9473	.9717	.9900	.9988	1.0000
10	.8780	.9036	.9294	.9544	.9763	.9922	.9992
11	.8639	.8877	.9122	.9366	.9599	.9798	.9937
12	.8505	.8727	.8958	.9193	.9425	.9643	.9825
14	.8257	.8452	.8656	.8867	.9085	.9303	.9515
15	.8142	.8325	.8517	.8717	.8924	.9136	.9347
16	.8033	.8205	.8386	.8575	.8772	.8974	.9181
18	.7828	.7983	.8144	.8313	.8490	.8674	.8864
20	.7642	.7781	.7926	.8078	.8237	.8404	.8577
21	.7554	.7686	.7825	.7969	.8121	.8279	.8443
24	.7311	.7427	.7547	.7672	.7803	.7940	.8082
25	.7236	.7347	.7462	.7582	.7707	.7837	.7973
27	.7095	.7196	.7302	.7412	.7527	.7646	.7770
28	.7027	.7125	.7227	.7332	.7442	.7557	.7676
30	.6899	.6990	.7084	.7182	.7283	.7389	.7498
32	..6779	.6864	.6951	.7042	.7136	.7233	.7335
33	.6722	.6804	.6888	.6976	.7066	.7160	.7258
35	.6613	.6689	.6768	.6850	.6934	.7022	.7112
36	.6560	.6634	.6711	.6790	.6871	.6956	.7043
40	.6365	.6430	.6498	.6567	.6639	.6713	.6790
42	.6274	.6336	.6400	.6466	.6533	.6603	.6675
45	.6147	.6204	.6263	.6323	.6386	.6449	.6515
48	.6028	.6081	.6136	.6192	.6249	.6308	.6368
49	.5991	.6042	.6095	.6150	.6206	.6263	.6322
50	.5954	.6004	.6056	.6109	.6164	.6220	.6277
55	.5780	.5825	.5872	.5919	.5967	.6017	.6067
60	.5624	.5664	.5706	.5748	.5791	.5836	.5881
70	.5350	.5384	.5418	.5453	.5489	.5525	.5562
80	.5119	.5147	.5176	.5206	.5235	.5266	.5297
90	.4919	.4943	.4968	.4993	.5019	.5045	.5072
100	.4743	.4765	.4787	.4809	.4831	.4854	.4877

TABLE 6.4.1. Contd.

(α = 0.05)

n|k	8	9	10	11	12	13	14	15
10	1.0000							
11	.9994	1.0000						
12	.9948	.9995	1.0000					
14	.9708	.9864	.9963	.9997	1.0000			
15	.9550	.9733	.9879	.9968	.9998	1.0000		
16	.9385	.9581	.9755	.9891	.9972	.9998	1.0000	
18	.9059	.9256	.9449	.9631	.9789	.9909	.9978	.9999
20	.8756	.8941	.9129	.9317	.9500	.9670	.9816	.9923
21	.8615	.8792	.8974	.9159	.9343	.9522	.9687	.9827
24	.8231	.8386	.8547	.8713	.8885	.9059	.9236	.9410
25	.8115	.8263	.8417	.8577	.8742	.8911	.9084	.9258
27	.7900	.8035	.8176	.8322	.8474	.8631	.8793	.8960
28	.7800	.7929	.8064	.8204	.8349	.8500	.8656	.8817
30	.7612	.7731	.7854	.7982	.8116	.8255	.8399	.8548
32	.7440	.7549	.7663	.7781	.7903	.8031	.8164	.8301
33	.7359	.7464	.7573	.7686	.7804	.7927	.8054	.8186
35	.7206	.7303	.7404	.7509	.7618	.7731	.7848	.7970
36	.7134	.7227	.7325	.7426	.7530	.7639	.7752	.7869
40	.6869	.6951	.7035	.7122	.7213	.7306	.7403	.7504
42	.6749	.6826	.6905	.6986	.7071	.7158	.7248	.7342
45	.6583	.6653	.6725	.6799	.6875	.6954	.7036	.7120
48	.6431	.6495	.6561	.6628	.6698	.6770	.6845	.6921
49	.6383	.6445	.6509	.6575	.6643	.6713	.6785	.6859
50	.6336	.6397	.6459	.6523	.6589	.6657	.6727	.6799
55	.6119	.6173	.6227	.6284	.6341	.6400	.6461	.6523
60	.5927	.5975	.6023	.6073	.6124	.6176	.6230	.6284
70	.5600	.5638	.5677	.5717	.5758	.5800	.5843	.5887
80	.5328	.5361	.5393	.5426	.5460	.5495	.5530	.5566
90	.5099	.5126	.5154	.5182	.5211	.5240	.5269	.5300
100	.4900	.4924	.4948	.4972	.4997	.5022	.5048	.5073

TABLE 6.4.1. Contd.

(α = 0.01)

n|k	1	2	3	4	5	6	7
5	.9859	.9980	1.0000				
6	.9741	.9911	.9990	1 0000			
7	.9608	.9804	.9939	.9994	1.0000		
8	.9470	.9676	.9845	.9955	.9996	1.0000	
9	.9332	.9538	.9725	.9874	.9966	.9998	1.0000
10	.9195	.9397	.9590	.9761	.9894	.9973	.9998
11	.9062	.9258	.9451	.9632	.9790	.9910	.9978
12	.8933	.9122	.9311	.9495	.9666	.9813	.9922
14	.8689	.8862	.9039	.9217	.9394	.9563	.9718
15	.8574	.8739	.8909	.9082	.9256	.9427	.9590
16	.8464	.8621	.8784	.8951	.9120	.9290	.9457
18	.8256	.8400	.8549	.8703	.8861	.9023	.9186
20	.8064	.8196	.8333	.8474	.8621	.8771	.8926
21	.7973	.8100	.8231	.8367	.8507	.8652	.8802
24	.7721	.7833	.7949	.8069	.8194	.8323	.8456
25	.7643	.7751	.7862	.7978	.8097	.8221	.8350
27	.7494	.7594	.7698	.7805	.7916	.8030	.8149
28	.7424	.7520	.7620	.7723	.7830	.7940	.8055
30	.7289	.7379	.7472	.7568	.7667	.7770	.7876
32	.7162	.7246	.7333	.7423	.7516	.7611	.7710
33	.7102	.7183	.7267	.7354	.7444	.7536	.7632
35	.6986	.7063	.7141	.7223	.7307	.7393	.7483
36	.6931	.7005	.7081	.7160	.7241	.7325	.7412
40	.6723	.6789	.6857	.6927	.6999	.7073	.7150
42	.6627	.6690	.6754	.6820	.6888	.6958	.7030
45	.6492	.6550	.6609	.6670	.6733	.6797	.6864
48	.6365	.6419	.6474	.6531	.6589	.6649	.6710
49	.6325	.6378	.6432	.6487	.6544	.6602	.6661
50	.6286	.6337	.6390	.6444	.6499	.6556	.6614
55	.6101	.6147	.6194	.6242	.6291	.6342	.6393
60	.5934	.5975	.6018	.6061	.6105	.6151	.6197
70	.5642	.5676	.5711	.5747	.5784	.5821	.5859
80	.5394	.5423	.5453	.5484	.5515	.5546	.5578
90	.5180	.5206	.5231	.5258	.5284	.5311	.5338
100	.4993	.5015	.5038	.5061	.5084	.5107	.5131

TABLE 6.4.1. Contd.

(α = 0.01)

n|k	8	9	10	11	12	13	14	15
10	1.0000							
11	.9998	1.0000						
12	.9982	.9999	1.0000					
14	.9847	.9939	.9987	.9999	1.0000			
15	.9738	.9860	.9946	.9989	1.0000	1.0000		
16	.9614	.9756	.9871	.9951	.9990	1.0000	1.0000	
18	.9348	.9506	.9654	.9785	.9889	.9960	.9993	1.0000
20	.9083	.9240	.9396	.9546	.9686	.9808	.9903	.9966
21	.8954	.9109	.9264	.9417	.9564	.9700	.9817	.9909
24	.8594	.8735	.8880	.9027	.9176	.9325	.9470	.9608
25	.8482	.8619	.8759	.8903	.9049	.9196	.9342	.9485
27	.8272	.8399	.8530	.8665	.8803	.8944	.9087	.9231
28	.8173	.8295	.8422	.8552	.8686	.8823	.8963	.9105
30	.7986	.8100	.8217	.8339	.8464	.8593	.8725	.8860
32	.7813	.7918	.8028	.8141	.8258	.8378	.8502	.8630
33	.7731	.7833	.7938	.8047	.8160	.8276	.8396	.8520
35	.7575	.7670	.7769	.7870	.7976	.8084	.8196	.8312
36	.7501	.7593	.7688	.7787	.7888	.7993	.8101	.8213
40	.7229	.7310	.7393	.7479	.7568	.7660	.7755	.7852
42	.7105	.7181	.7259	.7340	.7424	.7510	.7599	.7690
45	.6932	.7002	.7074	.7148	.7224	.7302	.7383	.7467
48	.6773	.6837	.6903	.6972	.7041	.7113	.7187	.7264
49	.6723	.6785	.6850	.6916	.6984	.7054	.7126	.7200
50	.6674	.6735	.6798	.6862	.6928	.6996	.7066	.7138
55	.6446	.6500	.6556	.6613	.6671	.6731	.6792	.6855
60	.6244	.6292	.6342	.6392	.6444	.6497	.6551	.6606
70	.5898	.5937	.5977	.6018	.6060	.6103	.6146	.6191
80	.5610	.5643	.5677	.5711	.5746	.5781	.5817	.5854
90	.5366	.5394	.5423	.5452	.5482	.5512	.5542	.5573
100	.5155	.5180	.5205	.5230	.5255	.5281	.5307	.5334

TABLE 6.1.2. Comparison of nominal and simulated critical values for T_{N3}.

n	$\alpha = 0.05$		$\alpha = 0.01$	
	Nominal	Tabulated	Nominal	Tabulated
10	3.829	3.82	4.099	4.00
20	5.344	5.30	5.639	5.60
50	6.996	6.82	7.386	7.34
100	8.048	7.77	8.473	8.27

TABLE 6.5.1. The probability P_a and Q_a for Murphy's test for n = 20 and α = 0.05.

θ	$\theta_1=\theta_2=\theta_3=\theta$			$\theta_1=\theta_2=\theta, \theta_3=0$	$\theta_1=\theta, \theta_2=\theta_3=0$
	P_a	P_a simulated	Q_a simulated	P_a	P_a
0	0.000	0.000	0.048	0.000	0.0000
1	0.004	0.004	0.051	0.001	0.0002
2	0.087	0.090	0.181	0.008	0.0006
3	0.452	0.460	0.538	0.030	0.0010
4	0.866	0.879	0.903	0.074	0.0016
5	0.991	0.991	0.991	0.132	0.0015
6	1.000	1.000	1.000	0.199	0.0012
7	1.000	1.000	1.000	0.266	0.0008
8	1.000	1.000	1.000	0.331	0.0005

TABLE 6.5.2. The probability P_a and Q_a for Murphy's test for n = 50 and α=0.05.

θ	$\theta_1=\theta_2=\theta_3=\theta$			$\theta_1=\theta_2=\theta, \theta_3=0$	$\theta_1=\theta, \theta_2=\theta_3=0$
	P_a	P_a simulated	Q_a simulated	P_a	P_a
0	0.000	0.000	0.032	0.000	0.0000
1	0.001	0.000	0.041	0.000	0.0000
2	0.060	0.056	0.193	0.004	0.0001
3	0.476	0.503	0.618	0.035	0.0005
4	0.928	0.924	0.953	0.152	0.0022
5	0.999	0.997	0.998	0.381	0.0044
6	1.000	1.000	1.000	0.650	0.0077
7	1.000	1.000	1.000	0.850	0.0120
8	1.000	1.000	1.000	0.951	0.0170

BIBLIOGRAPHY

BIBLIOGRAPHY

1. Anscombe, F.J. (1960). "Rejection of outliers". Technometrics, 2, 123-147.

2. Anscombe, F.J. (1961). "Examination of residuals". Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, University of California Press, Berkeley and Los Angeles, 1-36.

3. Anscombe, F.J. and Tukey, J.W. (1963). "The examination and analysis of residuals". Technometrics 5, 141-160.

4. Barnett, V. and Lewis, T. (1978). "Outliers in Statistical Data". John Wiley, New York.

5. Beckman, R.J. and Cook, R.D. (1983). "Outlier...s". Technometrics, 25, 119-149. (Response 161-163).

6. Bradu, D. and Hawkins, D.M. (1982). "Location of multiple outliers in two-way tables, using tetrads". Technometrics, 24, 103-108.

7. Bremner, J.M. (1978). "Algorithm AS 123. Mixtures of Beta Distributions". Applied Statistics, 27, 104-109.

8. Bross, I.D.J. (1961). "Outliers in patterned experiments : a strategic appraisal". Technometrics, 3, 91-102.

9. Cook, R.D. and Prescott, P. (1981). "On the accuracy of Bonferroni significance levels for detecting outliers in linear models". Technometrics, 23, 59-63.

10. Daniel, C. (1960). "Locating outliers in factorial experiments". Technometrics, 2, 149-156.

11. David, H.A. (1956). "On the application to statistics of an elementary theorem in probability". Biometrika, 43, 85-91.

12. David, H.A. (1981). "Order Statistics", 2nd ed., John Wiley, New York.

13. David, H.A., Hartley, H.O. and Pearson, E.S. (1954). "The distribution of the ratio, in a single normal sample, of range to standard deviation". Biometrika, 41, 482-493.

14. David, H.A. and Paulson, A.S. (1965). "The performance of several tests for outliers". Biometrika, 52, 429-436.

15. Dixon, W.J. (1953). "Processing data for outliers". Biometrics, 9, 74-89.

16. Doornbos, R. (1980). "Testing for an outlier in a linear model". Memorandum COSOR M80-01, Dept.of Mathematics, Eindhoven University of Technology, Eindhoven.

17. Doornbos, R. (1981). "Testing for a single outlier in a linear model". Biometrics, 37, 705-711.

18. Draper, N.R. and John, J.A. (1980). "Testing for three or fewer outliers in two-way tables". Technometrics, 22, 9-15.

19. Ellenberg, J.H. (1973). "The joint distribution of the standardized least squares residuals from a general linear regression". J.Amer.Statist.Ass., 68,941-943.

20. Ellenberg, J.H. (1976). "Testing for a single outlier from a general linear regression". Biometrics, 32, 637-645.

21. Galpin, J.S. and Hawkins, D.M. (1981). "Rejection of a single outlier in two-or three-way layouts". Technometrics, 23, 65-70.

22. Gentle, J.E. (1978). "Testing for outliers in linear regression". In Contributions to Survey Sampling and Applied Statistics, Papers in Honor of H.O. Hartley, Ed. H.A. David,p.223-233, Academic Press, New York.

23. Gentleman, J.F. (1980). "Finding the K most likely outliers in two-way tables". Technometrics, 22, 591-600.

24. Gentleman, J.F. and Wilk, M.B. (1975a). "Detecting outliers in a two-way table : I. Statistical behaviour of residuals". Technometrics, 17, 1-14.

25. Gentleman, J.F. and Wilk, M.B. (1975b). "Detecting outliers: II Supplementing the direct analysis of residuals". Biometrics, 31, 387-410.

26. Grubbs, F.E. (1950). "Sample criteria for testing outlying observations". Ann. Math. Statist., 21, 27-58.

27. Grubbs, F.E. and Beck, G. (1972). "Extension of sample sizes and percentage points for significance tests of outlying observations". Technometrics, 14, 847-854.

28. Hawkins, D.M. (1978). "Analysis of three tests for one or two outliers". Statistica Neerlandica, 32, 137-148.

29. Hawkins, D.M. (1980). "Identification of Outliers". Chapman and Hall, New York.

30. John, J.A. and Draper, N.R. (1978). "On testing for two outliers or one outlier in two-way tables". Technometrics, 20, 69-78.

31. Johnson, N.L. and Kotz, S. (1970). "Distributions in Statistics : Continuous Univariate Distributions", Vol. II. Houghton Mifflin Company, Boston.

32. Joshi, P.C. (1972). "Some slippage tests of mean for a single outlier in linear regression". Biometrika, 59, 109-120.

33. Joshi, P.C. (1975). "Some distribution theory results for a regression model". Ann. Inst.Statist. Math., 27, 309-317.

34. Kale, B.K. (1979). "Outliers - A Review". Journal of the Indian Statistical Association, 17, 51-67.

35. Lund, R.E. (1975). "Tables for an approximate test for outliers in linear models". Technometrics, 17, 473-476.

36. Margolin, B.H. (1977). "The distribution of internally studentized statistics via Laplace transform inversion". Biometrika, 64, 573-582.

37. McMillan, R.G. (1971). "Tests for one or two outliers in normal samples with unknown variance". Technometrics, 13, 87-100.

38. McMillan, R.G. and David, H.A. (1971). "Tests for one of two outliers in normal samples with known variance". Technometrics, 13, 75-85.

39. Mickey, M.R., Dunn, O.J. and Clark, V. (1967). "Note on the use of stepwise regression in detecting outliers". Computers and Biomedical Research 1, 105-111.

40. Moran, M.A. and McMillan, R.G. (1973). "Tests for one or two outliers in normal samples with unknown variance : a correction". Technometrics, 15, 637-640.

41. Murphy, R.B. (1951). "On Tests for Outlying Observations". Ph.D. Thesis, Princeton University.

42. Patnaik, P.B. (1949). "The non-central χ^2-and F-distributions and their applications". Biometrika, 36, 202-232.

43. Pearson, E.S.and Chandra Sekar, C. (1936). "The efficiency of statistical tools and a criterion for the rejection of outlying observations". Biometrika, 28, 308-320.

44. Pearson, E.S. and Hartley, H.O. (1970). "Biometrika Tables for Statisticians". Vol. I, 3rd ed., reprinted. Cambridge University Press, Cambridge.

45. Pearson, E.S. and Stephens, M.A. (1964). "The ratio of range to standard deviation in the same normal sample". Biometrika, 51, 484-487.

46. Pearson, K. (1968). "Tables of the Incomplete Beta-Function", 2nd ed. (with new Introduction by E.S. Pearson, and N.L. Johnson). Cambridge University Press, Cambridge.

47. Prescott, P. (1975). "An approximate test for outliers in linear models". Technometrics, 17, 129-132.

48. Quesenberry, C.P. and David, H.A. (1961). "Some tests for outliers". Biometrika, 48, 379-390.

49. Rao, C.R. (1973). "Linear Statistical Inference and Its Applications". 2nd ed., John Wiley, New York.

50. Rosner, B. (1975). "On the detection of many outliers". Technometrics, 17, 221-227.

51. Selby, S.M. and Girling, B. (Eds.) (1965). "Standard Mathematical Tables". 14th ed. The Chemical Rubber Co., Ohio.

52. Shapiro, S.S. and Wilk, M.B. (1965). "An analysis of variance test for normality (complete samples)". Biometrika, 52, 591-611.

53. Shapiro, S.S., Wilk, M.B. and Chen, H.J. (1968). "A comparative study of various tests for normality". J.Amer. Statist. Ass., 63, 1343-1372.

54. Srikantan, K.S. (1961). "Testing for the single outlier in a regression model". Sankhya A, 23,251-260.

55. Stefansky, W. (1971). "Rejecting outliers by maximum normed residual". Ann. Math. Statist., 42, 35-45.

56. Tietjen, G.L. and Moore, R.H. (1972). "Some Grubbs-type statistics for the detection of several outliers". Technometrics, 14,583-597.

APPENDIX I

SOLUTION OF $I_h(p,1/2) = \gamma$

Here we give an algorithm for numerical evaluation of the incomplete beta function $I_h(p,1/2)$, where p is a multiple of half. Then using this we develop an algorithm for inverse of incomplete beta function by numerical iteration procedure for small values of γ, $0 < \gamma < 1$. The calculations are performed on DEC 1090 Computer system.

a. Numerical evaluation of incomplete beta function

The incomplete beta function $I_h(p,q)$, is calculated using the recursive formula (Bremner, 1978)

$$(A.1) \qquad I_h(p,q) = I_h(p-1,q) - h^{p-1}(1-h)^q/[(p+q-1)\ B(p,q)] ,$$

where $p > 1$, $q > 0$ and $0 \leq h < 1$. For our purposes, we need $I_h(p,q)$ for $q = 1/2$ and p in multiples of half.

The recursion formula is started with

$$I_h(1/2,1/2) = \int_0^h t^{-1/2}\ (1-t)^{-1/2}\ dt/B(1/2,1/2)$$

$$= (2/\pi)\ \mathrm{Sin}^{-1}\ (h^{1/2}), \text{ and}$$

$$I_h(1,1/2) = \int_0^h (1-t)^{-1/2}\ dt/B(1,1/2)$$

$$= 1-(1-h)^{1/2}.$$

Equation (A.1) along with these initial values give $I_h(p,1/2)$ recursively for all values of p which are multiples of half.

b. Evaluation of inverse of incomplete beta function

The solution of equation

$$\text{(A.2)} \qquad \frac{1}{B(p,1/2)} \int_0^x t^{p-1} (1-t)^{-1/2} \, dt = \gamma ,$$

where p is a multiple of half and $\gamma > 0$ is small, is obtained by numerical iteration. For an initial value of x, we partially integrate (A.2) to get

$$\text{(A.3)} \qquad \frac{x^p (1-x)^{-1/2}}{p \, B(p,1/2)} - \frac{1}{2pB(p,1/2)} \int_0^x t^p (1-t)^{-3/2} \, dt = \gamma .$$

For small values of γ, the solution x of equation (A.2) is close to zero. Neglecting the second term on the L.H.S. of equation (A.3) and approximating $(1-x)^{-1/2}$ to 1, we get

$$x^p \cong \gamma . p . B(p,1/2) .$$

This gives an initial value

$$\text{(A.4)} \qquad x_0 = [\gamma . p . B(p,1/2)]^{1/p} .$$

In general x_0 gives an "over estimate" of true value and requires a larger number of iterations. A slightly better approximation is obtained by neglecting the second term of equation (A.3) and considering

$$\frac{x^p (1-x)^{-1/2}}{p \, B(p,1/2)} = \gamma .$$

On raising it to power 1/p, we have

$$x = [\gamma . p . B(p,1/2)]^{1/p} \, (1-x)^{1/2p} .$$

Using equation (A.4), and substituting x_o for x in the factor $(1-x)^{1/2p}$, we get

$$x \simeq x_o(1-x_o)^{1/2p}.$$

Our calculations show that, in general, this value of x is an under estimate. But

(A.5) $$x_1 = x_o(1-x_o/2p)$$

gives best approximate value, even for small values of p.

This is used for starting the iteration procedure.

The final solution is obtained by Newton-Raphson method using the relation

(A.6) $$x_i = x_{i-1} - f(x_{i-1})/f'(x_{i-1}),$$

where

$$f(x) = I_x(p,1/2) - \gamma .$$

The iteration is terminated when the absolute difference between two successive iterations is not greater than the required accuracy.

The function subprogram XBETA calculates the incomplete beta function $I_h(p,1/2)$ for a given value of p and h. For this subprogram, the complete beta functions are supplied from the main calling program. Thus, this program can calculate the incomplete beta function for any value of h in (0,1) and p, a multiple of half with $p \leq 100$.

The function subprogram XINBTA calculates the inverse of incomplete beta function for a given value of p and γ. The error is indicated by IR which is initially assigned the value zero. It is equal to 1 if the iteration procedure does not converge in 20 steps, and is equal to 2 if the solution at any stage is greater than 1. The final solution is V. It is equal to x_{n+1} if the iteration converges in n steps, otherwise it is equal to x_{21}. For checking the convergence of the iteration, the accuracy ACU is taken to be equal to 10^{-6}.

LANGUAGE

Fortran 10

STRUCTURE

SUBROUTINE XINBTA(CB,P,PROB,V,IR)

Formal parameters

CB	Real input vector of length 200	:	This vector specifies the complete beta functions B(p,1/2) for p = 0.5(0.5)100.
P	Real input	:	This is the first parameter 'p' of the incomplete beta function.
PROB	Real input	:	This is the desired probability γ.
V	Real output	:	This is the final value given by the iteration procedure. (In the ith step, this is the value equal to x_{i+1} $(i = 1,2,\ldots,n)$. Let n be the number of iterations required for convergence to achieve desired accuracy, then V = x_{n+1}, otherwise V is x_{21}).

IR Integer output : Error indicator
= 1 if the iteration does not converge
= 2 if $|x_i| \geq 1$ for some i = 1,2,...,n.
= 0 otherwise.

Auxiliary Algorithm

Subroutine XINBTA calls subroutine XBETA.

ACCURACY

The program XINBTA gives accurate values upto 5 decimal places. The accuracy can be increased upto 15 decimal places by assigning the value ACU, appearing in the program, equal to the necessary accuracy requirement. But in that case the number of iterations will have to be suitably relaxed.

```
C
      SUBROUTINE XINBTA(CB,P,PROB,V,IR)
C
C     THIS SUBROUTINE CALCULATES THE INVERSE OF INCOMPLETE
C     BETA FUNCTION,WHEN THE SECOND PARAMETER OF THE BETA
C     FUNCTION IS HALF AND THE FIRST PARAMETER IS A MULTIPLE
C     OF HALF
C
      EXTERNAL XBETA,DSQRT,DABS,DEXP,DLOG,DATAN
      DOUBLE PRECISION X(50),CB(200),F,F1,XBETA,P,B1,BETA,H,
     1PROB,ACU,V,P1
C
C     N IS THE NUMBER OF ITERATIONS
C
      N=20
C
C     INITIALIZE CONSTANTS
C
      ACU=0.000001
      IR=0
      P1=P+P
      IP1=P1
      B1=CB(IP1)
C
C     CALCULATION OF INITIAL APPROXIMATION
C
```

```
      X(1)=DEXP((DLOG(PROB*P*B1))/P)
      X(1)=X(1)*(1-X(1)/P1)
      I=1
C
C     SOLVE FOR V USING NEWTON-RAPHSON METHOD,USING THE FUNCTION
C     XBETA
C
C     1/F1 DENOTES THE DERIVATIVE OF THE FUNCTION F
C
 3    H=DABS(X(I))
      IF(H.GE.1)GO TO 31
      F=XBETA(CB,P,H)-PROB
      F1=DEXP(DLOG(B1)-(P-1)*DLOG(X(I))+(DLOG(1-X(I)))/2)
      X(I+1)=X(I)-F*F1
      IF(DABS(X(I+1)-X(I))/X(I).LE.ACU)GO TO 20
      IF(I.GT.N)GO TO 30
      I=I+1
      GO TO 3
 20   V=X(I+1)
      GO TO 10
C
C     THE FOLLOWING DEFINES THE ERROR INDICATOR IR
C
 30   IR=1
      GO TO 10
 31   IR=2
```

```
 10   RETURN
      END
C
C
      DOUBLE PRECISION FUNCTION XBETA(CB,P,H)
C
C     THIS PROGRAM CALCULATES THE INCOMPLETE BETA FUNCTION
C     UPTO THE POINT H WITH PARAMETERS P AND HALF
C
      EXTERNAL DSQRT,DATAN
      DOUBLE PRECISION IHBETA(5000),CB(200),H,P,PI,C,RH,IHB,
     1B1,P1,DASIN,P2,ACU,TERM
C
C     INITIALISING CONSTANTS
C
      P1=0
      PI=3.14159265358979 32
      ACU=0.1D-16
      I1=0
C
C     CALCULATION OF INITIAL VALUES FOR THE RECURRENCE RELATION
C
  8   I1=I1+1
      P1=P1+0.5
      P2=P1+P1+0.1
      IP2=P2
```

```
      B1=CB(IP2)
      RH=DSQRT(H)
      C=DSQRT(1-H)
      IF(P1.EQ.0.5)GO TO 5
      IF(P1.EQ.1.0)GO TO 6
      GO TO 20
 5    IF(RH.EQ.1)GO TO 7
      DASIN=DATAN(RH/DSQRT(1-RH*RH))
      GO TO 9
 7    DASIN=PI/2.
 9    IHBETA(I1)=2*DASIN/PI
      GO TO 10
 6    IHBETA(I1)=(1-C)
      GO TO 10
C
C     CALCULATION FOR HIGHER VALUES OF P USING RECURRENCE
C     RELATION
C
 20   IHB=IHBETA(I1-2)
      TERM=DSQRT((H**(2*P1-2))*(1-H))/((P1-0.5)*B1)
      IHBETA(I1)=IHB-TERM
      IF(TERM.LT.ACU)GO TO 17
 10   IF(P1.LT.P)GO TO 8
 17   XBETA=IHBETA(I1)
      RETURN
      END
```

```
C
C                CALLING PROGRAM
C                ---------------
C
C       CALCULATION OF INVERSE OF INCOMPLETE BETA
C       FUNCTION
C
        DOUBLE PRECISION CB(200),P,PROB,V,H
C       CALCULATION OF COMPLETE BETA FUNCTIONS
C
        CB(1)=3.14159265358979 32
        CB(2)=2.0
        DO 10 I=1,198
        P=DFLOAT(I)/2.
 10     CB(I+2)=(P/(P+0.5))*CB(I)
C
C       CALCULATION OF THE MAIN RESULT FOR A GIVEN
C       P AND PROB
C
        P=6.5
        PROB=0.0004928
        CALL XINBTA(CB,P,PROB,V,IR)
        IF(IR)21,20,21
 21     IF(IR-2)14,15,14
 14     PRINT 100
 100    FORMAT(//10X,'THE ITERATION DOES NOT CONVERGES')
```

```
      GO TO 500
15    PRINT 200
200   FORMAT(//10X,'H IS GREATER THAN 1')
      GO TO 500
20    PRINT 300,V
300   FORMAT(//10X,F8.5)
500   STOP
      END
```